

Information Structures for Feedback Capacity of Channels with Memory and Transmission Cost: Stochastic Optimal Control & Variational Equalities-Part I

Christos K. Kourtellaris and Charalambos D. Charalambous

Abstract

The Finite Transmission Feedback Information (FTFI) capacity is characterized for any class of channel conditional distributions $\{\mathbf{P}_{B_i|B^{i-1},A_i} : i = 0, 1, \dots, n\}$ and $\{\mathbf{P}_{B_i|B_{i-M}^{i-1},A_i} : i = 0, 1, \dots, n\}$, where M is the memory of the channel, $B^n \triangleq \{B_j : j = \dots, 0, 1, \dots, n\}$ are the channel outputs and $A^n \triangleq \{A_j : j = \dots, 0, 1, \dots, n\}$ are the channel inputs. The characterizations of FTFI capacity, are obtained by first identifying the information structures of the optimal channel input conditional distributions $\mathcal{P}_{[0,n]} \triangleq \{\mathbf{P}_{A_i|A^{i-1},B^{i-1}} : i = 0, \dots, n\}$, which maximize directed information

$$C_{A^n \rightarrow B^n}^{FB} \triangleq \sup_{\mathcal{P}_{[0,n]}} I(A^n \rightarrow B^n), \quad I(A^n \rightarrow B^n) \triangleq \sum_{i=0}^n I(A^i; B_i | B^{i-1}).$$

The main theorem states, for any channel with memory M , the optimal channel input conditional distributions occur in the subset satisfying conditional independence $\mathcal{P}_{[0,n]} \triangleq \{\mathbf{P}_{A_i|A^{i-1},B^{i-1}} = \mathbf{P}_{A_i|B_{i-M}^{i-1}} : i = 1, \dots, n\}$, and the characterization of FTFI capacity is given by

$$C_{A^n \rightarrow B^n}^{FB,M} \triangleq \sup_{\mathcal{P}_{[0,n]}} \sum_{i=0}^n I(A_i; B_i | B_{i-M}^{i-1}).$$

Similar conclusions are derived for problems with transmission cost constraints of the form $\frac{1}{n+1} \mathbf{E} \left\{ \sum_{i=0}^n \gamma_i(A_i, T^i B^{n-1}) \right\} \leq \kappa$, $\kappa > 0$, where $\{\gamma_i(A_i, T^i B^{n-1}) : i = 0, 1, \dots, n\}$ is any class of multi-letter functions such that $T^i B^{n-1} = \{B_{i-1}, B_{i-2}, \dots, B_{i-K}\}$ or $T^i B^{n-1} = \{B^{i-1}\}$, for $i = 0, \dots, n$ and K a nonnegative integer.

The methodology utilizes stochastic optimal control theory, to identify the control process, the controlled process, and a variational equality of directed information, to derive upper bounds on $I(A^n \rightarrow B^n)$, which are achievable over specific subsets of channel input conditional distributions $\mathcal{P}_{[0,n]}$, which are characterized by conditional independence. For channels with limited memory, this implies the transition probabilities of the channel output process are also of limited memory.

For any of the above classes of channel distributions and transmission cost functions, a direct analogy, in terms of conditional independence, of the characterizations of FTFI capacity and Shannon's capacity formulae of Memoryless Channels is identified.

I. INTRODUCTION

Feedback capacity of channel conditional distributions, $\{\mathbf{P}_{B_i|B^{i-1},A_i} : i = 0, 1, \dots, n\}$, where $a^n \triangleq \{\dots, a_{-1}, a_0, a_1, \dots, a_n\} \in \mathbb{A}^n$, $b^n \triangleq \{\dots, b_{i-1}, b_0, b_1, \dots, b_n\} \in \mathbb{B}^n$, are the channel input and output sequences, respectively, is often defined by maximizing directed information [1], [2], $I(A^n \rightarrow B^n)$, from channel input sequences $a^n \in \mathbb{A}^n$ to channel output sequences $b^n \in \mathbb{B}^n$, over an

admissible set of channel input conditional distributions, $\mathcal{P}_{[0,n]} \triangleq \{\mathbf{P}_{A_i|A^{i-1},B^{i-1}} : i = 0, 1, \dots, n\}$ (with feedback), as follows.

$$C_{A^\infty \rightarrow B^\infty}^{FB} \triangleq \liminf_{n \rightarrow \infty} \frac{1}{n+1} C_{A^n \rightarrow B^n}^{FB}, \quad C_{A^n \rightarrow B^n}^{FB} \triangleq \sup_{\mathcal{P}_{[0,n]}} I(A^n \rightarrow B^n), \quad (\text{I.1})$$

$$I(A^n \rightarrow B^n) \triangleq \sum_{i=0}^n I(A^i; B_i | B^{i-1}) = \sum_{i=0}^n \mathbf{E}_\mu \left\{ \log \left(\frac{d\mathbf{P}_{B_i|B^{i-1},A^i}(\cdot|B^{i-1},A^i)}{d\mathbf{P}_{B_i|B^{i-1}}(\cdot|B^{i-1})}(B_i) \right) \right\} \quad (\text{I.2})$$

where for each i , $\mathbf{P}_{B_i|B^{i-1}}$, is the conditional probability distribution of the channel output process, and $\mathbf{E}_\mu\{\cdot\}$ denotes expectation with respect to the joint distribution \mathbf{P}_{A^i,B^i} , for $i = 0, \dots, n$, for a fixed initial distribution¹ $\mathbf{P}_{A^{-1},B^{-1}} \equiv \mu(da^{-1},db^{-1})$ of the initial data $\{(A^{-1},B^{-1}) = (a^{-1},b^{-1})\}$. In (I.1), \liminf can be replaced by \lim if it exists and it is finite. For finite alphabet spaces sufficient conditions are identified in [3]. Moreover, for finite alphabet spaces \sup can be replaced by maximum, since probability mass functions can be viewed as closed and bounded subsets of finite-dimensional spaces. However, for countable or abstract alphabet spaces it is more difficult, and often requires an analysis using the topology of weak convergence of probability distributions, because information theoretic measures are not necessarily continuous with respect to pointwise convergence of probability distributions, and showing compactness of the set of distributions is quite involved.

It is shown in [3]–[5], using tools from [6]–[13], under appropriate conditions which include abstract alphabets, that the quantity $C_{A^\infty \rightarrow B^\infty}^{FB}$ is the supremum of all achievable rates of a sequence of feedback codes $\{(n, M_n, \epsilon_n) : n = 0, 1, \dots\}$, defined as follows.

(a) A set of uniformly distributed source messages $\mathcal{M}_n \triangleq \{1, \dots, M_n\}$ and a set of encoding strategies, mapping source messages into channel inputs of block length $(n+1)$, defined by

$$\mathcal{E}_{[0,n]}^{FB} \triangleq \left\{ g_i : \mathcal{M}_n \times \mathbb{A}^{i-1} \times \mathbb{B}^{i-1} \mapsto \mathbb{A}_i : i = 0, \dots, n : a_0 = g_0(w), a_1 = g_1(w, a_0, b_0), \dots, a_n = g_n(w, a^{n-1}, b^{n-1}), \quad w \in \mathcal{M}_n \right\}, \quad n = 0, 1, \dots \quad (\text{I.3})$$

The codeword for any $w \in \mathcal{M}_n$ is $u_w \in \mathbb{A}^n$, $u_w = (g_0(w), g_1(w, a_0, b_0), \dots, g_n(w, a^{n-1}, b^{n-1}))$, and $\mathcal{C}_n = (u_1, u_2, \dots, u_{M_n})$ is the code for the message set \mathcal{M}_n , and $\{A^{-1}, B^{-1}\} = \{\emptyset\}$. In general, the code may depend on the initial data, depending on the convention, i.e., $(A^{-1}, B^{-1}) = (a^{-1}, b^{-1})$, which are often known to the encoder and decoder.

(b) Decoder measurable mappings $d_{0,n} : \mathbb{B}^n \mapsto \mathcal{M}_n$, such that the average probability of decoding error satisfies²

$$\mathbf{P}_e^{(n)} \triangleq \frac{1}{M_n} \sum_{w \in \mathcal{M}_n} \mathbf{P}^g \left\{ d_{0,n}(B^n) \neq w | W = w \right\} \equiv \mathbf{P}^g \left\{ d_{0,n}(B^n) \neq W \right\} \leq \epsilon_n, \quad w \in \mathcal{M}_n$$

and the decoder may also assume knowledge of the initial data.

The coding rate or transmission rate is defined by $r_n \triangleq \frac{1}{n+1} \log M_n$. A rate R is said to be an achievable rate, if there exists a code sequence satisfying $\lim_{n \rightarrow \infty} \epsilon_n = 0$ and $\liminf_{n \rightarrow \infty} \frac{1}{n+1} \log M_n \geq R$. The feedback capacity is supremum of all achievable rates, i.e., defined by $C \triangleq \sup\{R : R \text{ is achievable}\}$.

The underlying assumption for $C_{A^\infty \rightarrow B^\infty}^{FB}$ to correspond to feedback capacity is that the source process $\{X_i : i = 0, \dots\}$ to be encoded and transmitted over the channel has finite entropy rate, and satisfies the following conditional independence [2].

$$\mathbf{P}_{B_i|B^{i-1},A^i,X^k} = \mathbf{P}_{B_i|B^{i-1},A^i} \quad \forall k \in \{0, 1, \dots, n\}, \quad i = 0, \dots, n \quad (\text{I.4})$$

Coding theorems for channels with memory with and without feedback are developed extensively over the years, in an anthology of papers, such as, [3]–[15], in three direction. For jointly stationary ergodic processes, for information stable processes, and for arbitrary nonstationary and nonergodic processes. Since many of the coding theorems presented in the above references are either directly applicable or applicable subject to the assumptions imposed in these references, the main emphasis of the current investigation is on the characterizations of FTFI capacity, for different channels with transmission cost.

¹The subscript notation on probability distributions and expectation, i.e., \mathbf{P}_μ and $\mathbf{E}_\mu\{\cdot\}$ is often omitted because it is clear from the context.

²The superscript on expectation, i.e., \mathbf{P}^g indicates the dependence of the distribution on the encoding strategies.

Similarly, feedback capacity with transmission cost is defined by

$$C_{A^\infty \rightarrow B^\infty}^{FB}(\kappa) \triangleq \liminf_{n \rightarrow \infty} \frac{1}{n+1} C_{A^n \rightarrow B^n}^{FB}(\kappa), \quad C_{A^n \rightarrow B^n}^{FB}(\kappa) \triangleq \sup_{\mathcal{P}_{[0,n]}(\kappa)} I(A^n \rightarrow B^n) \quad (I.5)$$

$$\mathcal{P}_{[0,n]}(\kappa) \triangleq \left\{ \mathbf{P}_{A_i|A^{i-1}, B^{i-1}}, i = 1, \dots, n : \frac{1}{n+1} \mathbf{E} \left(\sum_{i=0}^n \gamma_i(T^i A^n, T^i B^{n-1}) \right) \leq \kappa \right\} \quad (I.6)$$

where for each i , $T^i a^n \subseteq \{\dots, a_{-1}, a_0, a_1, \dots, a_i\}$, $T^i b^{n-1} \subseteq \{\dots, b_{-1}, b_0, b_1, \dots, b_{i-1}\}$, for $i = 0, \dots, n$, and these are either fixed or nondecreasing with i , for $i = 0, 1, \dots, n$.

The hardness of such extremum problems of capacity, and in general, of other similar problems of information theory, is attributed to the form of the directed information density or sample pay-off functional, defined by

$$I_{A^n \rightarrow B^n}(a^n, b^n) \triangleq \sum_{i=0}^n \log \left(\frac{d\mathbf{P}_{B_i|B^{i-1}, A^i}(\cdot|b^{i-1}, a^i)}(b_i)}{d\mathbf{P}_{B_i|B^{i-1}}(\cdot|b^{i-1})}(b_i) \right) \quad (I.7)$$

which is not fixed. Rather, the pay-off $I_{A^n \rightarrow B^n}(a^n, b^n)$ depends on the channel output conditional probabilities $\{\mathbf{P}_{B_i|B^{i-1}}(db_i|b^{i-1}) : i = 0, \dots, n\}$, which in turn depends on the channel input conditional distributions $\{\mathbf{P}_{A_i|A^{i-1}, B^{i-1}}(da_i|a^{i-1}, b^{i-1}) : i = 0, \dots, n\} \in \mathcal{P}_{[0,n]}(\kappa)$, chosen to maximize the expectation $\mathbf{E}\{I_{A^n \rightarrow B^n}(A^n, B^n)\}$. This means, given a specific channel conditional distribution and a transmission cost function, the information structure of the channel input conditional distribution denoted by $\mathcal{I}_i^P \subseteq \{a^{i-1}, b^{i-1}\}, i = 0, \dots, n$, which maximizes directed information (i.e., the dependence of the optimal channel input conditional distribution on past information), needs to be identified, and then used to obtain the characterizations of the Finite Transmission Feedback Information (FTFI) capacity, $C_{A^n \rightarrow B^n}^{FB}(\kappa)$, and feedback capacity $C_{A^\infty \rightarrow B^\infty}^{FB}(\kappa)$.

For memoryless stationary channels (such as, Discrete Memoryless Channels (DMCs)), described by $\mathbf{P}_{B_i|B^{i-1}, A^i} = \mathbf{P}_{B_i|A_i} \equiv \mathbf{P}_{B|A}, i = 0, 1, \dots, n$, without feedback (with transmission cost constraints if necessary), Shannon [16] characterized channel capacity by the well-known two-letter formulae

$$C \triangleq \max_{\mathbf{P}_A} I(A; B) = \max_{\mathbf{P}_A} \mathbf{E} \left\{ \log \left(\frac{d\mathbf{P}_{A,B}(\cdot, \cdot)}{d(\mathbf{P}_A(\cdot) \times \mathbf{P}_B(\cdot))}(A, B) \right) \right\} \quad (I.8)$$

where $\mathbf{P}_{A,B}(da, db) = \mathbf{P}_{B|A}(db|a) \otimes \mathbf{P}_A(da)$ is the joint distribution, $\mathbf{P}_{B|A}(db|a)$ is the channel conditional distribution, $\mathbf{P}_A(da)$ is the channel input distribution, $\mathbf{P}_B(db) = \int \mathbf{P}_{B|A}(db|a) \otimes \mathbf{P}_A(da)$ is the channel output distribution, and $\mathbf{E}\{\cdot\}$ denotes expectation with respect to $\mathbf{P}_{A,B}$.

This characterization is often obtained by identifying the information structures of optimal channel input distributions, via the upper bound

$$C_{A^n; B^n} \triangleq \max_{\mathbf{P}_{A^n}} I(A^n; B^n) \leq \max_{\mathbf{P}_{A_i}, i=0, \dots, n} \sum_{i=0}^n I(A_i; B_i) \leq (n+1)C \quad (I.9)$$

since this bound is achievable, when the channel input distribution satisfies conditional independence $\mathbf{P}_{A_i|A^{i-1}}(da_i|a^{i-1}) = \mathbf{P}_{A_i}(da_i), i = 0, 1, \dots, n$, and moreover C is obtained, when $\{A_i : i = 0, 1, \dots, n\}$ is identically distributed, which then implies the joint process $\{(A_i, B_i) : i = 0, 1, \dots, n\}$ is independent and identically distributed, and $I(A^n; B^n) = (n+1)I(A; B)$.

For memoryless stationary channels with feedback, the characterization of feedback capacity, denoted by C^{FB} , is shown by Shannon and subsequently Dobrushin [17] to correspond to the capacity without feedback, i.e., $C^{FB} = C$. This fundamental formulae is often shown by first applying the converse to the coding theorem, to show that feedback does not increase capacity (see [18] for discussion on this subtle issue), which then implies

$$\mathbf{P}_{A_i|A^{i-1}, B^{i-1}}(da_i|a^{i-1}, b^{i-1}) = \mathbf{P}_{A_i}(da_i), \quad i = 0, 1, \dots, n \quad (I.10)$$

and $C^{FB} = C$ is obtained if $\{A_i : i = 0, 1, \dots, n\}$ is identically distributed. That is, since feedback does not increase capacity, then mutual information and directed information are identical, in view of (I.10). However, as pointed out elegantly by Massey [2], for channels with feedback it will be a mistake to use the arguments in (I.9) to derive C^{FB} . The conditional independence condition (I.10) implies that the *Information Structure* of the maximizing channel input distributions is the *Null Set*.

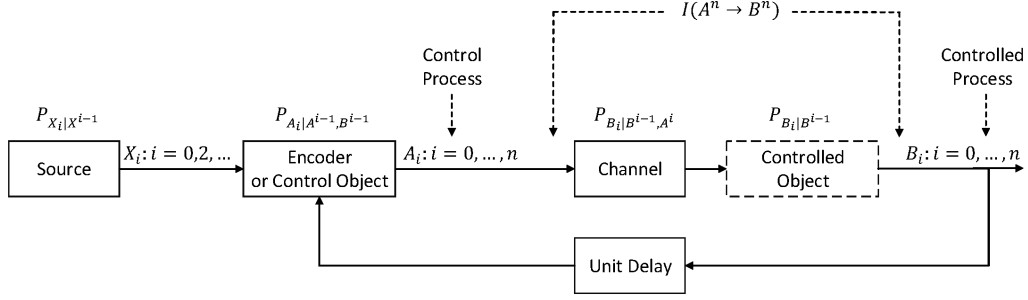


Fig. I.1. Communication block diagram and its analogy to stochastic optimal control.

The methodology developed in this paper, establishes a direct analogy between the conditional independence properties (I.10) of capacity achieving channel input distributions of memoryless channels and corresponding properties for channels with memory and feedback. To this date, no such systematic methodology is developed in the literature, to determine the information structure of optimal channel input distributions, which maximize directed information $I(A^n \rightarrow B^n)$, via achievable upper bounds over subsets of channel input conditional distributions $\overline{\mathcal{P}}_{[0,n]} \subseteq \mathcal{P}_{[0,n]}(\kappa)$, which satisfy conditional independence, and to characterize the corresponding FTFI capacity and feedback capacity.

In this first part, of a two-part investigation, the main objective is to

develop a methodology to identify information structures of optimal channel input conditional distributions, for channels with memory, with and without transmission cost, of extremum problems defined by (I.1) and (I.5), and to characterize the corresponding FTFI capacity and feedback capacity.

This is addressed by utilizing connections between stochastic optimal control and information theoretic concepts, as follows. The theory of stochastic optimal control is linked to the identification of information structures of optimal channel input conditional distributions and to the characterization of FTFI capacity, by first establishing the the following analogy.

The information measure $I(A^n \rightarrow B^n)$ is the pay-off;
the channel output process $\{B_i : i = 0, 1, \dots, n\}$ is the controlled process;
the channel input process $\{A_i : i = 0, 1, \dots, n\}$ is the control process.

Indeed, as depicted in Fig.I, the channel output process $\{B_i : i = 0, 1, \dots, n\}$ is controlled, by controlling its conditional probability distribution $\{\mathbf{P}_{B_i|B^{i-1}}(db_i|b^{i-1}) : i = 0, \dots, n\}$ via the choice of the transition probability distribution $\{\mathbf{P}_{A_i|A^{i-1}, B^{i-1}}(da_i|a^{i-1}, b^{i-1}) : i = 0, \dots, n\} \in \mathcal{P}_{[0,n]}$ called the control object.

As in any stochastic optimal control problem, given a channel distribution, the distribution of the initial data, and a transmission cost function, the main objective is to determine the controlled object conditional distribution, the control object conditional distribution, and the functional dependence of the pay-off on these objects.

However, unlike classical stochastic optimal control theory, the directed information density pay-off (i.e., (I.7)), depends nonlinearly on the channel output conditional distributions $\{\mathbf{P}_{B_i|B^{i-1}}(db_i|b^{i-1}) : i = 0, \dots, n\}$, induced by the control objects, a variational equality is linked to tight upper bounds on directed information $I(A^n \rightarrow B^n)$, which together with the stochastic optimal control analogy, are shown to be achievable over specific subsets of the control objects. These achievable bounds depend on the structural properties of the channel conditional distributions and the transmission cost functions.

The methodology is based on a two-step procedure, as follows. Given a class of channel conditional distributions, the distribution of the initial data, and a class of transmission cost functions, any candidate of the optimal channel input conditional distribution

or control object, which maximizes $I(A^n \rightarrow B^n)$ is shown to satisfy the following conditional independence.

$$\mathbf{P}_{A_i|A^{i-1}, B^{i-1}}(da_i|a^{i-1}, b^{i-1}) = \mathbf{P}(da_i|\mathcal{I}_i^{\mathbf{P}^*}) \equiv \mathbb{P}\{A_i \in da_i|\mathcal{I}_i^{\mathbf{P}^*}\}, \quad \mathcal{I}_i^{\mathbf{P}^*} \subseteq \{a^{i-1}, b^{i-1}\}, \quad i = 0, 1, \dots, n, \quad (\text{I.11})$$

$$\mathcal{I}_i^{\mathbf{P}^*} \triangleq \text{Information Structure of optimal channel input distributions which maximizes } I(A^n \rightarrow B^n) \text{ for } i = 0, 1, \dots, n. \quad (\text{I.12})$$

Moreover, the information structure $\mathcal{I}_i^{\mathbf{P}^*}, i = 0, 1, \dots, n$, is specified by the memory of the channel conditional distribution, and the dependence of the transmission cost function on the channel input and output symbols. Consequently, the dependence of the joint distribution of $\{(A_i, B_i) : i = 0, \dots, n\}$, the conditional distribution of the channel output process $\{B_i : i = 0, \dots, n\}$, i.e., $\{\mathbf{P}_{B_i|B^{i-1}}(db_i|b^{i-1}) : i = 0, \dots, n\}$ and the directed information density $\iota_{A^n \rightarrow B^n}(A^n, B^n)$, on the control object, is determined, and the characterization of FTFI capacity is obtained.

The characterization of feedback capacity is obtained from the per unit time limiting version of the characterization of the FTFI capacity.

These structural properties of channel input distribution, which maximize directed information settle various open problems in Shannon's information theory, which include the role of feedback signals to control, via the control process (channel input), the controlled process (channel output process), and the design of encoders which achieve the characterizations of FTFI capacity and capacity.

Indeed, in the second part of this two-part investigation [19], and based on these structural properties, a methodology is developed to realize optimal channel input conditional distributions, by information lossless randomized strategies (driven by uniform Random Variables on $[0, 1]$, which can generate any distribution), and to construct encoders, which achieve the characterizations of FTFI capacity and feedback capacity. Applications of the results of this first part, to various channel models, which include Multiple-Input Multiple Output (MIMO) Gaussian Channel Models with memory, are found in the second part of this investigation. In this part, we give an illustrative simple example to clarify the importance of information structures of optimal channel input distributions, in reduction of computation complexity, and to indicate the analogy to Shannon's two-letter capacity formulae of DMCs.

A. Literature Review of Feedback Capacity of Channels with Memory

Although, in this paper we do not treat channels with memory dependence on past channel input symbols, for completeness we review such literature, and we discuss possibly extensions of our methodology to such channels at the end of the paper.

Cover and Pombra [18] (see also Ihara [11]) characterized the feedback capacity of non-stationary Additive Gaussian Noise (AGN) channels with memory, defined by

$$B_i = A_i + Z_i, \quad i = 0, 1, \dots, n, \quad \frac{1}{n+1} \sum_{i=0}^n \mathbf{E}\{|A_i|^2\} \leq \kappa, \quad \kappa \in [0, \infty) \quad (\text{I.13})$$

where $\{Z_i : i = 0, 1, \dots, n\}$ is a real-valued (scalar) jointly non-stationary Gaussian process, denoted by $N(\mu_{Z^n}, K_{Z^n})$, and "Aⁿ is causally related to Zⁿ" defined by³ $\mathbf{P}_{A^n, Z^n}(da^n, dz^n) = \otimes_{i=0}^n \mathbf{P}_{A_i|A^{i-1}, Z^{i-1}}(da_i|a^{i-1}, z^{i-1}) \otimes \mathbf{P}_{Z^n}(dz^n)$. The authors in [18] characterized the capacity of this non-stationary AGN channel, by first characterizing the FTFI capacity formulae⁴ via the expression

$$C_{0,n}^{FB,CP}(\kappa) \triangleq \max_{\left\{(\Gamma, K_{V^n}) : \frac{1}{n+1} \mathbf{E}\{tr(A^n(A^n)^T)\} \leq \kappa, \quad A^n = \Gamma Z^n + V^n\right\}} H(B^n) - H(Z^n) \quad (\text{I.14})$$

where $V^n \triangleq \{V_i : i = 0, 1, \dots, n\}$ is a Gaussian process $N(0, K_{V^n})$, orthogonal to $Z^n \triangleq \{Z_i : i = 0, \dots, n\}$, and Γ is lower diagonal time-varying matrix with deterministic entries. The feedback capacity is given by [18] $C^{FB,CP}(\kappa) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n+1} C_{0,n}^{FB,CP}(\kappa)$.

³ [18], page 39, above Lemma 5.

⁴ The methodology in [18] utilizes the converse coding theorem to obtain an upper bound on the entropy $H(B^n)$, by restricting $\{A_i : i = 0, \dots, n\}$ to a Gaussian process.

Kim [20] revisited the stationary version of feedback capacity characterization of the Cover and Pombra AGN channel, and utilized frequency domain methods, and their relations to scalar Riccati equations, and showed that if the noise power spectral density corresponds to a stationary Gaussian autoregressive moving-average model of order K , then a K -dimensional generalization of the Schalkwijk-Kailath [21] coding scheme achieves feedback capacity. Yang, Kavcic, and Tatikonda [22] analyzed the feedback capacity of stationary AGN channels, re-visited the Cover and Pombra AGN channel, and proposed solution methods based on dynamic programming, to perform the optimization in (I.14). Butman [23], [24] evaluated the performance of linear feedback schemes for AGN channels, when the noise is described by an autoregressive moving average model. A historical account regarding Gaussian channels with memory and feedback, related to the the Cover and Pombra [18] AGN channel, is found in [20].

Recently, for finite alphabet channels with memory and feedback, expressions of feedback capacity are derived for the trapdoor channel by Permuter, Cuff, Van Roy and Tsachy [25], for the Ising Channel by Elishco and Permuter [26], for the Post(a, b) channel by Permuter, Asnani and Tsachy [27], all without transmission cost constraints, and in [28] for the BSSC(α, β) with and without feedback and transmission cost. Tatikonda, Yang and Kavcic [29] showed that if the input to the channel and the channel state are related by a one-to-one mapping, and the channel assumes a specific structure, specifically, $\{\mathbf{P}_{B_i|A_i, A_{i-1}} : i = 0, \dots, n\}$, then dynamic programming can be used to compute the feedback capacity expression given in [29]. Chen and Berger [30] analyzed the Unit Memory Channel Output (UMCO) channel $\{\mathbf{P}_{B_i|B_{i-1}, A_i} : i = 0, \dots, n\}$, under the assumption that the optimal channel input distribution is $\{\mathbf{P}_{A_i|B_{i-1}} : i = 0, \dots, n\}$. The authors in [30] showed that the UMCO channel can be transformed to one with state information, and that under certain conditions on the channel and channel input distributions, dynamic programming can be used to compute feedback capacity.

B. Discussion of Main Results and Methodology

In this paper, the emphasis is on any combination of the following classes of channel distributions and transmission cost functions⁵.

Channel Distributions

$$\text{Class A. } \mathbf{P}_{B_i|B^{i-1}, A_i}(db_i|b^{i-1}, a^i) = \mathbf{P}_{B_i|B^{i-1}, A_i}(db_i|b^{i-1}, a_i), \quad i = 0, \dots, n. \quad (\text{I.15})$$

$$\text{Class B. } \mathbf{P}_{B_i|B^{i-1}, A_i}(db_i|b^{i-1}, a^i) = \mathbf{P}_{B_i|B_{i-M}^{i-1}, A_i}(db_i|b_{i-M}^{i-1}, a_i), \quad i = 0, \dots, n. \quad (\text{I.16})$$

Transmission Costs

$$\text{Class A. } \gamma_i(T^i a^n, T^i b^{n-1}) = \gamma_i^A(a_i, b^{i-1}), \quad i = 0, \dots, n, \quad (\text{I.17})$$

$$\text{Class B. } \gamma_i(T^i a^n, T^i b^{n-1}) = \gamma_i^B(a_i, b_{i-K}^{i-1}), \quad i = 0, \dots, n. \quad (\text{I.18})$$

Here, $\{K, M\}$ are nonnegative finite integers and the following convention is used.

$$\text{If } M = 0 \text{ then } \mathbf{P}_{B_i|B_{i-M}^{i-1}, A_i}(db_i|b_{i-M}^{i-1}, a_i) \Big|_{M=0} \equiv \mathbf{P}_{B_i|A_i}(db_i|a_i), \quad \text{for } i = 0, 1, \dots, n.$$

$$\text{If } K = 0 \text{ then } \gamma_i^B(a_i, b_{i-K}^{i-1}) \Big|_{K=0} \equiv \gamma_i^B(a_i), \quad i = 0, \dots, n.$$

Thus, for $M = 0$ the above convention implies the channel degenerates to the memoryless channel $\mathbf{P}_{B_i|A_i}(db_i|a_i), i = 0, 1, \dots, n$.

The above classes of channel conditional distributions may be induced by various nonlinear channel models (NCM), such as, nonlinear and linear time-varying Autoregressive models, and nonlinear and linear channel models expressed in state space form [31]. Such classes are investigated in [19].

An over view of the methodology and results obtained, is discussed below, to illustrate analogies to Shannon's two-letter capacity formulae (I.8) and conditional independence conditions (I.10) in relation to (I.11).

⁵The methodology developed in the paper can be extended to channels and transmission cost functions with past dependence on channel input symbols; however, such generalizations are beyond the scope of this paper.

1) **Channels of Class A and Transmission Cost of Class A or B:** In Theorem III.1, Step 1 of a two-step procedure, based on stochastic optimal control, is applied to channel distributions of Class A, $\{\mathbf{P}_{B_i|B^{i-1},A_i}(db_i|b^{i-1},a_i) : i = 0, 1, \dots, n\}$, to show the optimal channel input conditional distribution, which maximizes $I(A^n \rightarrow B^n)$ satisfies conditional independence $\mathbf{P}_{A_i|A^{i-1},B^{i-1}} = \mathbf{P}_{A_i|B^{i-1}}, i = 0, \dots, n$, and hence it occurs in the subset

$$\overline{\mathcal{P}}_{[0,n]}^A \triangleq \{\mathbf{P}_{A_i|B^{i-1}}(da_i|b^{i-1}) : i = 0, \dots, n\} \subset \mathcal{P}_{[0,n]}. \quad (\text{I.19})$$

This means that for each i , the information structures of the maximizing channel input distribution is $\mathcal{S}_i^{\mathbf{P}} \triangleq \{b^{i-1}\} \subset \{a^{i-1}, b^{i-1}\}$, for $i = 0, 1, \dots, n$.

The characterization of the FTFI capacity is

$$C_{A^n \rightarrow B^n}^{FBA} = \sup_{\overline{\mathcal{P}}_{[0,n]}^A} \sum_{i=0}^n I(A_i; B_i | B^{i-1}). \quad (\text{I.20})$$

If a transmission cost $\mathcal{P}_{[0,n]}(\kappa)$ is imposed corresponding to any of the functions $\gamma_i^A(a_i, b^{i-1})$, $\gamma_i^B(a_i, b_{i-K}^{i-1}), i = 0, 1, \dots, n$, the characterization of the FTFI capacity is

$$C_{A^n \rightarrow B^n}^{FBA}(\kappa) \triangleq \sup_{\overline{\mathcal{P}}_{[0,n]}^A \cap \mathcal{P}_{[0,n]}(\kappa)} \sum_{i=0}^n I(A_i; B_i | B^{i-1}). \quad (\text{I.21})$$

2) **Channels of Class B Transmission Cost of Class A or B:** In Theorem III.3, Step 2 of the two-step procedure, a variational equality of directed information, is applied to channel distributions of Class B, $\{\mathbf{P}_{B_i|B_{i-M}^{i-1},A_i}(db_i|b_{i-M}^{i-1},a_i) : i = 0, 1, \dots, n\}$, to show the optimal channel input conditional distribution, which maximizes $I(A^n \rightarrow B^n)$ satisfies conditional independence $\mathbf{P}_{A_i|A^{i-1},B^{i-1}} = \mathbf{P}_{A_i|B_{i-M}^{i-1}}, i = 0, \dots, n$, and hence it occurs in the subset

$$\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,M} \triangleq \{\mathbf{P}_{A_i|B_{i-M}^{i-1}}(da_i|b_{i-M}^{i-1}) : i = 0, 1, \dots, n\}. \quad (\text{I.22})$$

The characterization of the FTFI capacity is then given by the following expression.

$$C_{A^n \rightarrow B^n}^{FBB,M} = \sup_{\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,M}} \sum_{i=0}^n I(A_i; B_i | B_{i-M}^{i-1}). \quad (\text{I.23})$$

If a transmission cost $\mathcal{P}_{[0,n]}(\kappa)$ is imposed corresponding to cost functions of Class B, $\{\gamma_i^B(a_i, b_{i-K}^{i-1}) : i = 0, \dots, n\}$, the optimal channel input conditional distribution occurs in the subset $\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,J} \cap \mathcal{P}_{[0,n]}(\kappa)$, where $J \triangleq \max\{M, K\}$.

The characterization of the FTFI capacity is given by the following expression.

$$C_{A^n \rightarrow B^n}^{FBB,J}(\kappa) = \sup_{\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,J} \cap \mathcal{P}_{[0,n]}(\kappa)} \sum_{i=0}^n \int \log \left(\frac{d\mathbf{P}_{B_i|B_{i-M}^{i-1},A_i}(\cdot|b_{i-M}^{i-1},a_i)}{d\mathbf{P}_{B_i|B_{i-J}^{i-1}}(\cdot|b_{i-J}^{i-1})}(b_i) \right) \mathbf{P}_{B_{i-J}^i,A_i}(db_{i-J},da_i), \quad J \triangleq \max\{M, K\}. \quad (\text{I.24})$$

where

$$\mathbf{P}_{B_{i-J}^i,A_i}(db_{i-J},da_i) = \mathbf{P}_{B_i|B_{i-M}^{i-1},A_i}(db_i|b_{i-M}^{i-1},a_i) \otimes \mathbf{P}_{A_i|B_{i-J}^{i-1}}(da_i|b_{i-J}^{i-1}) \otimes \mathbf{P}_{B_{i-J}^{i-1}}(db_{i-J}^{i-1}), \quad i = 0, 1, \dots, n, \quad (\text{I.25})$$

$$\mathbf{P}_{B_i|B_{i-J}^{i-1}}(db_i|b_{i-J}^{i-1}) = \int \mathbf{P}_{B_i|B_{i-M}^{i-1},A_i}(db_i|b_{i-M}^{i-1},a_i) \otimes \mathbf{P}_{A_i|B_{i-J}^{i-1}}(da_i|b_{i-J}^{i-1}), \quad i = 0, 1, \dots, n. \quad (\text{I.26})$$

The above expressions imply the channel output process or controlled process $\{B_i : i = 0, \dots, n\}$ is a J -order Markov process. On the other hand, if a transmission cost $\mathcal{P}_{[0,n]}(\kappa)$ is imposed corresponding to $\gamma_i^A(a_i, b^{i-1}), i = 0, 1, \dots, n$, the optimal channel input distribution occurs in the set $\overline{\mathcal{P}}_{[0,n]}^A \cap \mathcal{P}_{[0,n]}(\kappa)$.

The above characterizations of FTFI capacity (and by extension of feedback capacity characterizations) state that the information structure of the optimal channel input conditional distribution is determined by $\max\{M, K\}$, where M specifies the order of the memory of the channel conditional distribution, and K specifies the dependence of the transmission cost function, on past

channel output symbols.

These structural properties of optimal channel input conditional distributions are analogous to those of memoryless channels, and they hold for finite, countable and abstract alphabet spaces (i.e., continuous), and channels defined by nonlinear models, state space models, autoregressive models, etc.

The following special cases illustrate the explicit analogy to Shannon's two-letter capacity formulae of memoryless channels.

Special Case- $M = 2, K = 1$. For any channel $\{\mathbf{P}_{B_i|B_{i-1}, B_{i-2}, A_i}(db_i|b_{i-1}, b_{i-2}, a_i) : i = 0, 1, \dots, n\}$, and transmission cost function $\{\gamma_i^{B,1}(a_i, b_{i-1}), i = 1, \dots, n\}$, from (I.24)-(I.26), the optimal channel input conditional distribution occurs in the subset

$$\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2}(\kappa) \triangleq \left\{ \mathbf{P}_{A_i|B_{i-1}, B_{i-2}}(da_i|b_{i-1}, b_{i-2}), i = 0, 1, \dots, n : \frac{1}{n+1} \mathbf{E} \left\{ \sum_{i=0}^n \gamma_i^{B,1}(A_i, B_{i-1}) \right\} \leq \kappa \right\}. \quad (\text{I.27})$$

The information structure of the optimal channel input conditional distribution implies the joint distribution of (A_i, B_i) , conditioned on (A^{i-1}, B^{i-1}) , is given by

$$\mathbf{P}_{A_i, B_i|A^{i-1}, B^{i-1}} = \mathbf{P}_{A_i, B_i|A_{i-1}, B_{i-1}, B_{i-2}} \equiv \mathbf{P}_{B_i|A_i, B_{i-1}, B_{i-2}} \otimes \mathbf{P}_{A_i|B_{i-1}, B_{i-2}}, \quad i = 0, 1, \dots, n. \quad (\text{I.28})$$

the channel output process $\{B_i : i = 0, \dots, n\}$ is a second-order Markov process, i.e.,

$$\mathbf{P}_{B_i|B^{i-1}} = \mathbf{P}_{B_i|B_{i-1}, B_{i-2}} = \int_{\mathbb{A}_i} \mathbf{P}_{B_i|B_{i-1}, B_{i-2}, A_i}(db_i|b_{i-1}, b_{i-2}, a_i) \otimes \mathbf{P}_{A_i|B_{i-1}, B_{i-2}}(da_i|b_{i-1}, b_{i-2}), \quad i = 0, 1, \dots, n. \quad (\text{I.29})$$

and that the characterization of the FTFI capacity is given by the following 4-letter expression at each time $i = 0, \dots, n$.

$$C_{A^n \rightarrow B^n}^{FB, B, 2}(\kappa) \triangleq \sup_{\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2}(\kappa)} \sum_{i=0}^n I(A_i; B_i|B_{i-1}, B_{i-2}). \quad (\text{I.30})$$

$$= \sup_{\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2}(\kappa)} \sum_{i=0}^n \mathbf{E} \left\{ \ell_i(A_i, S_i) \right\}, \quad S_j \triangleq (B_{j-1}, B_{j-2}) \quad j = 0, \dots, n, \quad (\text{I.31})$$

where the pay-off $\ell_j(\cdot, \cdot)$ is given by

$$\ell_j(a_j, s_j) \triangleq \int_{\mathbb{B}_j} \log \left(\frac{d\mathbf{P}_{B_j|S_j, A_j}(\cdot|s_j, a_j)}{d\mathbf{P}_{B_j|S_j}(\cdot|s_j)}(b_j) \right) \mathbf{P}_{B_j|S_j, A_j}(db_j|s_j, a_j), \quad j = 0, \dots, n \quad (\text{I.32})$$

Moreover, if the channel input distribution is restricted to a time-invariant distribution, i.e., $\mathbf{P}_{A_i|S_i}(da_i|s) \equiv \mathbf{P}^\infty(da_i|s), i = 0, \dots$, and the channel distribution is time-invariant, then the transition probability distribution of $\{S_i : i = 0, \dots\}$ is time-invariant, and $\mathbf{P}_{B_i|S_{i-1}} \equiv \mathbf{P}^\infty(db_i|s), i = 0, \dots$, is also time-invariant. Consequently, the per unit limiting version of (I.30), specifically, $C_{A^\infty \rightarrow B^\infty}^{FB, B, 2}(\kappa)$, under conditions which ensure ergodicity, is characterized by time-invariant and the ergodic distribution of $\{S_i : i = 0, \dots, n\}$ [32].

Special Case- $M = 2, K = 0$. This means no transmission cost is imposed, and hence the supremum in (I.30) is over $\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2} \triangleq \{\mathbf{P}_{A_i|B_{i-1}, B_{i-2}}(da_i|b_{i-1}, b_{i-2}), i = 0, 1, \dots, n\}$. Let $C_t^{B,2} : \mathbb{B}_{t-1} \times \mathbb{B}_{t-2} \mapsto \mathbb{R}$ denote the cost-to-go corresponding to (I.31), with $K = 0$, from time "t" to the terminal time "n" given the values of the output $S_t = (B_{t-1}, B_{t-2}) = (b_{t-1}, b_{t-2})$.

Then the cost-to-go satisfies the following dynamic programming recursions.

$$C_n^{B,2}(s_n) = \sup_{\mathbf{P}_{A_n|S_n}} \left\{ \int_{\mathbb{A}_n \times \mathbb{B}_n} \log \left(\frac{d\mathbf{P}_{B_n|S_n, A_n}(\cdot|s_n, a_n)}{d\mathbf{P}_{B_n|S_n}(\cdot|s_n)}(b_n) \right) \mathbf{P}_{B_n|S_n, A_n}(db_n|s_n, a_n) \otimes \mathbf{P}_{A_n|S_n}(da_n|s_n) \right\}, \quad (\text{I.33})$$

$$\begin{aligned} C_t^{B,2}(s_t) &= \sup_{\mathbf{P}_{A_t|S_t}} \left\{ \int_{\mathbb{A}_t \times \mathbb{B}_t} \log \left(\frac{d\mathbf{P}_{B_t|S_t, A_t}(\cdot|s_t, a_t)}{d\mathbf{P}_{B_t|S_t}(\cdot|s_t)}(b_t) \right) \mathbf{P}_{B_t|S_t, A_t}(db_t|s_t, a_t) \otimes \mathbf{P}_{A_t|S_t}(da_t|s_t) \right. \\ &\quad \left. + \int_{\mathbb{A}_t \times \mathbb{B}_t} C_{t+1}^{B,2}(s_t) \mathbf{P}_{B_t|S_t, A_t}(db_t|s_t, a_t) \otimes \mathbf{P}_{A_t|S_t}(da_t|s_t) \right\}, \quad t = n-1, n-2, \dots, 0. \end{aligned} \quad (\text{I.34})$$

The characterization of the FTFI capacity and feedback capacity are expressed via the $C_0^{B,2}(s_0)$ and the fixed distribution

$\mu_{B_{-1}, B_{-2}}(db_{-1}, db_{-2})$ by

$$C_{A^n \rightarrow B^n}^{FB, B, 2} = \int_{\mathbb{B}_{-1} \times \mathbb{B}_{-2}} C_0^{B, 2}(s_0) \mu_{B_{-1}, B_{-2}}(ds_0), \quad C_{A^\infty \rightarrow B^\infty}^{FB, B, 2} \triangleq \liminf_{n \rightarrow \infty} \frac{1}{n+1} C_{A^n \rightarrow B^n}^{FB, B, 2}. \quad (I.35)$$

Obviously, even for finite “ n ”, from the above recursions, we deduce that the information structure, $\{S_t = B_{t-1}, B_{t-2} : t = 0, \dots, n\}$, of the control object, namely, $\{\mathbf{P}_{A_t|S_t} : t = 0, \dots, n\}$, induces conditional probabilities $\{\mathbf{P}_{B_t|S^t} = \mathbf{P}_{B_t|S_t} : t = 0, \dots, n\}$ which are 2nd order Markov, i.e., $\{\mathbf{P}_{S_{t+1}|S^t} = \mathbf{P}_{S_{t+1}|S_t} : t = 0, \dots, n-1\}$, resulting in a significant reduction in computational complexity of the above dynamic programming recursions. Clearly, for any fixed $S_0 = s_0$, then $C_{A^\infty \rightarrow B^\infty}$ depends on the initial state $S_0 = s_0$. However, if the channel is time-invariant and the distributions $\{\mathbf{P}_{A_t|S_t} : t = 0, \dots, \}$ are either restricted or converge to time-invariant distributions, and the corresponding transition probabilities $\{\mathbf{P}_{S_{t+1}|S^t} = \mathbf{P}_{S_{t+1}|S_t} : t = 0, \dots, n-1\}$ are irreducible and aperiodic, then there is unique invariant distribution for $\{S_i : i = 0, \dots, \}$ and $C_{A^\infty \rightarrow B^\infty}^{B, 2}$ is independent of the initial distribution $\mu_{B_{-1}, B_{-2}}(ds_0)$. Such questions are addressed in [30] for the channel $\{\mathbf{P}_{B_i|B_{i-1}, A_i}(db_i|b_{i-1}, a_i) : i = 0, 1, \dots, n\}$. They can be addressed from the general theory of per unit time-infinite horizon Markov decision theory [33], and more generally by solving explicitly the above dynamic programming recursions and investigating their per unit-time limits (see [34]).

Special Case- $M = K = 1$. If the channel is the so-called Unit Memory Channel Output (UMCO) defined by $\{\mathbf{P}_{B_i|B_{i-1}, A_i}(db_i|b_{i-1}, a_i) : i = 0, 1, \dots, n\}$, and the transmission cost function is $\{\gamma_i^{B, 1}(a_i, b_{i-1}), i = 1, \dots, n\}$, the optimal channel input conditional distribution occurs in the subset

$$\overset{\circ}{\mathcal{P}}_{[0, n]}^{B, 1}(\kappa) \triangleq \left\{ \mathbf{P}_{A_i|B_{i-1}}(da_i|b_{i-1}), i = 0, 1, \dots, n : \frac{1}{n+1} \mathbf{E} \left\{ \sum_{i=0}^n \gamma_i^{B, 1}(A_i, B_{i-1}) \right\} \leq \kappa \right\}. \quad (I.36)$$

and the characterization of the FTFI capacity degenerates to the following sums of a 3-letter expressions.

$$C_{A^n \rightarrow B^n}^{FB, B, 1}(\kappa) \triangleq \sup_{\overset{\circ}{\mathcal{P}}_{[0, n]}^{B, 1}(\kappa)} \sum_{i=0}^n I(A_i; B_i | B_{i-1}). \quad (I.37)$$

The importance of variational equalities to identify information structures of capacity achieving channel input conditional distributions is first applied in [35]. For the BSSC(α, β) (which is a special case of the UMCO) with transmission cost, it is shown in [28], that the characterizations of feedback capacity and capacity without feedback, admit closed form expressions. Moreover, this channel is matched to the Binary Symmetric Markov Source through the use of nonanticipative Rate Distortion Function (RDF) in [36]. That is, there is a perfect duality between the BSSC(α, β) with transmission cost and the Binary Symmetric Markov Source with a single letter distortion function.

Recently, the results of this paper are applied in [34] (see also [37]) to derive sequential necessary and sufficient conditions to optimize the characterizations of FTFI capacity. Moreover, using the necessary and sufficient conditions closed form expressions for the optimal channel input distributions and feedback capacity, are obtained for various applications examples defined on finite alphabet spaces. This paper includes in Section IV, an illustrative example, which reveals many silent properties of capacity achieving distributions, with and without feedback, for a simple first-order Gaussian Linear Channel Model.

A detailed investigation of the characterization of FTFI capacity, and feedback capacity, of Multiple-Input Multiple Output (MIMO) Gaussian Linear Channel Models with memory is included in the second part of this two-part investigation [19].

II. DIRECTED INFORMATION AND DEFINITIONS OF EXTREMUM PROBLEMS OF CAPACITY

In this section, the notation adopted in the rest of the paper is introduced, and a variational equality of directed information is recalled from [38].

The following notation is used throughout the paper.

\mathbb{Z} : set of integer;

\mathbb{N}_0 : set of nonnegative integers $\{0, 1, 2, \dots\}$;

$(\Omega, \mathcal{F}, \mathbb{P})$: probability space, where \mathcal{F} is the σ -algebra generated by subsets of Ω ;

$\mathcal{B}(\mathbb{W})$: Borel σ -algebra of a given topological space \mathbb{W} ;

$\mathcal{M}(\mathbb{W})$: set of all probability measures on $\mathcal{B}(\mathbb{W})$ of a Borel space \mathbb{W} ;

$\mathcal{K}(\mathbb{V}|\mathbb{W})$: set of all stochastic kernels on $(\mathbb{V}, \mathcal{B}(\mathbb{V}))$ given $(\mathbb{W}, \mathcal{B}(\mathbb{W}))$ of Borel spaces \mathbb{W}, \mathbb{V} .

All spaces (unless stated otherwise) are complete separable metric spaces, also called Polish spaces, i.e., Borel spaces. This generalization is judged necessary to treat simultaneously discrete, finite alphabet, real-valued \mathbb{R}^k or complex-valued \mathbb{C}^k random processes for any positive integer k , etc.

A. Basic Notions of Probability

The product measurable space of the two measurable spaces $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$ is denoted by $(\mathbb{X} \times \mathbb{Y}, \mathcal{B}(\mathbb{X}) \odot \mathcal{B}(\mathbb{Y}))$, where $\mathcal{B}(\mathbb{X}) \odot \mathcal{B}(\mathbb{Y})$ is the product σ -algebra generated by $\{A \times B : A \in \mathcal{B}(\mathbb{X}), B \in \mathcal{B}(\mathbb{Y})\}$.

A Random Variable (RV) defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ by the mapping $X : (\Omega, \mathcal{F}) \mapsto (\mathbb{X}, \mathcal{B}(\mathbb{X}))$ induces a probability measure $\mathbf{P}(\cdot) \equiv \mathbf{P}_X(\cdot)$ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ as follows⁶.

$$\mathbf{P}(A) \equiv \mathbf{P}_X(A) \triangleq \mathbb{P}\{\omega \in \Omega : X(\omega) \in A\}, \quad \forall A \in \mathcal{B}(\mathbb{X}). \quad (\text{II.38})$$

A RV is called discrete if there exists a countable set $\mathcal{S}_X \triangleq \{x_i : i \in \mathbb{N}\}$ such that $\sum_{x_i \in \mathcal{S}_X} \mathbb{P}\{\omega \in \Omega : X(\omega) = x_i\} = 1$. The probability measure $\mathbf{P}_X(\cdot)$ is then concentrated on points in \mathcal{S}_X , and it is defined by

$$\mathbf{P}_X(A) \triangleq \sum_{x_i \in \mathcal{S}_X \cap A} \mathbb{P}\{\omega \in \Omega : X(\omega) = x_i\}, \quad \forall A \in \mathcal{B}(\mathbb{X}). \quad (\text{II.39})$$

If the cardinality of \mathcal{S}_X is finite then the RV is finite-valued and it is called a finite alphabet RV.

Given another RV $Y : (\Omega, \mathcal{F}) \mapsto (\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$, for each Borel subset B of \mathbb{Y} and any sub-sigma-field $\mathcal{G} \in \mathcal{F}$ (collection of events) the conditional probability of event $\{Y \in B\}$ given \mathcal{G} is defined by $\mathbb{P}\{Y \in B | \mathcal{G}\}(\omega)$, and this is an \mathcal{G} -measurable function $\forall \omega \in \Omega$. This conditional probability induces a conditional probability measure on $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$ defined by $\mathbf{P}(B | \mathcal{G})(\omega)$, which is a version of $\mathbb{P}\{Y \in B | \mathcal{G}\}(\omega)$. For example, if \mathcal{G} is the σ -algebra generated by RV X , and $B = dy$, then $\mathbf{P}_{Y|X}(dy|X)(\omega)$ is called the conditional distribution of RV Y given RV X . The conditional distribution of RV Y given $X = x$ is denoted by $\mathbf{P}_{Y|X}(dy|X = x) \equiv \mathbf{P}_{Y|X}(dy|x)$. Such conditional distributions are equivalently described by stochastic kernels or transition functions $\mathbf{K}(\cdot|\cdot)$ on $\mathcal{B}(\mathbb{Y}) \times \mathbb{X}$, mapping \mathbb{X} into $\mathcal{M}(\mathbb{Y})$ (the space of probability measures on $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$), i.e., $x \in \mathbb{X} \mapsto \mathbf{K}(\cdot|x) \in \mathcal{M}(\mathbb{Y})$, and hence the distributions are parametrized by $x \in \mathbb{X}$.

The family of probability measures on $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$ parametrized by $x \in \mathbb{X}$, is defined by

$$\mathcal{K}(\mathbb{Y}|\mathbb{X}) \triangleq \{\mathbf{K}(\cdot|x) \in \mathcal{M}(\mathbb{Y}) : x \in \mathbb{X} \text{ and } \forall F \in \mathcal{B}(\mathbb{Y}), \text{ the function } \mathbf{K}(F|\cdot) \text{ is } \mathcal{B}(\mathbb{X})\text{-measurable.}\}.$$

B. FTFI Capacity and Variational Equality

The channel input and channel output alphabets are sequences of measurable spaces $\{(\mathbb{A}_i, \mathcal{B}(\mathbb{A}_i)) : i \in \mathbb{Z}\}$ and $\{(\mathbb{B}_i, \mathcal{B}(\mathbb{B}_i)) : i \in \mathbb{Z}\}$, respectively, and their history spaces are the product spaces $\mathbb{A}^{\mathbb{Z}} \triangleq \times_{i \in \mathbb{Z}} \mathbb{A}_i$, $\mathbb{B}^{\mathbb{Z}} \triangleq \times_{i \in \mathbb{Z}} \mathbb{B}_i$. These spaces are endowed with their respective product topologies, and $\mathcal{B}(\Sigma^{\mathbb{Z}}) \triangleq \odot_{i \in \mathbb{Z}} \mathcal{B}(\Sigma_i)$ denotes the σ -algebra on $\Sigma^{\mathbb{Z}}$, where $\Sigma_i \in \{\mathbb{A}_i, \mathbb{B}_i\}$, $\Sigma^{\mathbb{Z}} \in \{\mathbb{A}^{\mathbb{N}}, \mathbb{B}^{\mathbb{Z}}\}$, generated by cylinder sets. Thus, for any $n \in \mathbb{Z}$, $\mathcal{B}(\Sigma^n)$ denote the σ -algebras of cylinder sets in $\Sigma^{\mathbb{Z}}$, with

⁶The subscript on X is often omitted.

bases over $C_i \in \mathcal{B}(\Sigma_i)$, $i = \dots, -1, 0, 1, \dots, n$, respectively. Points in $\mathbb{A}^n, \mathbb{B}^n$ are denoted by $a^n \triangleq \{\dots, a_{-1}, a_0, a_1, \dots, a_n\} \in \mathbb{A}^n$, $b^n \triangleq \{\dots, b_{-1}, b_0, b_1, \dots, b_n\} \in \mathbb{B}^n$. Similarly, points in $\mathbb{Z}_k^m \triangleq \times_{j=k}^m \mathbb{Z}_j$ are denoted by $z_k^m \triangleq \{z_k, z_{k+1}, \dots, z_m\} \in \mathbb{Z}_k^m$, $(k, m) \in \mathbb{Z} \times \mathbb{Z}$. We often restrict \mathbb{Z} to \mathbb{N}_0 .

Channel Distribution with Memory. A sequence of stochastic kernels or distributions defined by

$$\mathcal{C}_{[0,n]} \triangleq \left\{ Q_i(db_i|b^{i-1}, a^i) = \mathbf{P}_{B_i|B^{i-1}, A^i} \in \mathcal{K}(\mathbb{B}_i|\mathbb{B}^{i-1} \times \mathbb{A}^i) : i = 0, 1, \dots, n \right\}. \quad (\text{II.40})$$

At each time instant i the conditional distribution of channel output B_i is affected causally by previous channel output symbols $b^{i-1} \in \mathbb{B}^{i-1}$ and current and previous channel input symbols $a^i \in \mathbb{A}^i$, $i = 0, 1, \dots, n$.

Channel Input Distribution with Feedback. A sequence of stochastic kernels defined by

$$\mathcal{P}_{[0,n]} \triangleq \left\{ P_i(da_i|a^{i-1}, b^{i-1}) = \mathbf{P}_{A_i|A^{i-1}, B^{i-1}} \in \mathcal{K}(\mathbb{A}_i|\mathbb{A}^{i-1} \times \mathbb{B}^{i-1}) : i = 0, 1, \dots, n \right\}. \quad (\text{II.41})$$

At each time instant i the conditional distribution of channel input A_i is affected causally by past channel inputs and output symbols $\{a^{i-1}, b^{i-1}\} \in \mathbb{A}^{i-1} \times \mathbb{B}^{i-1}$, $i = 0, 1, \dots, n$. Hence, the information structure of the channel input distribution at time instant i is $\mathcal{I}_i^P \triangleq \{a^{i-1}, b^{i-1}\} \in \mathbb{A}^{i-1} \times \mathbb{B}^{i-1}$, $i = 0, 1, \dots, n$.

Admissible Histories. For each $i = -1, 0, \dots, n$, we introduce the space \mathbb{G}^i of admissible histories of channel input and output symbols, as follows. Define

$$\mathbb{G}^i \triangleq \mathbb{B}^{-1} \times \mathbb{A}_0 \times \mathbb{B}_0 \times \dots \times \mathbb{A}_{i-1} \times \mathbb{B}_{i-1} \times \mathbb{A}_i \times \mathbb{B}_i, \quad i = 0, \dots, n, \quad \mathbb{G}^{-1} = \mathbb{B}^{-1}. \quad (\text{II.42})$$

A typical element of \mathbb{G}^i is a sequence of the form $(b^{-1}, a_0, b_0, \dots, a_i, b_i)$. We equip the space \mathbb{G}^i with the natural σ -algebra $\mathcal{B}(\mathbb{G}^i)$, for $i = -1, 0, \dots, n$. Hence, for each i , the information structure of the channel input distribution is

$$\mathcal{I}_i^P \triangleq \left\{ B^{-1}, A_0, B_0, \dots, A_{i-1}, B_{i-1} \right\}, \quad i = 0, 1, \dots, n, \quad \mathcal{I}_0^P \triangleq \{B^{-1}\} \quad (\text{II.43})$$

This implies at time $i = 0$, the initial distribution is $P_0(da_0|a^{-1}, b^{-1}) = P_0(da_0|\mathcal{I}_0^P) = P_0(da_0|b^{-1})$. However, we can modify \mathcal{I}_0^P to consider an alternative convention such as $\mathcal{I}_0^P = \{\emptyset\}$ or $\mathcal{I}_0^P = \{a^{-1}, b^{-1}\}$.

Joint and Marginal Distributions. Given any channel input conditional distribution $\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, 1, \dots, n\} \in \mathcal{P}_{[0,n]}$, any channel distribution $\{Q_i(db_i|b^{i-1}, a^i) : i = 0, 1, \dots, n\} \in \mathcal{C}_{[0,n]}$, and the initial probability distribution $\mathbf{P}(db^{-1}) \equiv \mu(db^{-1}) \in \mathcal{M}(\mathbb{G}^{-1})$, the induced joint distribution $\mathbf{P}^P(da^n, db^n)$ on the canonical space $(\mathbb{G}^n, \mathcal{B}(\mathbb{G}^n))$ is defined uniquely, and a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ carrying the sequence of RVs $\{(A_i, B_i) : i = 0, \dots, n\}$ and B^{-1} can be constructed, as follows⁷.

$$\begin{aligned} \mathbb{P}\{A^n \in da^n, B^n \in db^n\} &\triangleq \mathbf{P}^P(db^{-1}, da_0, db_0, \dots, da_n, db_n), \quad n \in \mathbb{N}_0 \\ &= \mu(db^{-1}) \otimes P_0(da_0|b^{-1}) \otimes Q_0(db_0|b^{-1}, a_0) \otimes P_1(da_1|b^{-1}, b_0, a_0) \\ &\quad \otimes \dots \otimes Q_{n-1}(db_{n-1}|b^{n-2}, a^{n-1}) \otimes P_n(da_n|b^{n-1}, a^{n-1}) \otimes Q_n(db_n|b^{n-1}, a^n) \end{aligned} \quad (\text{II.44})$$

$$\equiv \mu(db^{-1}) \otimes_{j=0}^n \left(Q_j(db_j|b^{j-1}, a^j) \otimes P_j(da_j|a^{j-1}, b^{j-1}) \right). \quad (\text{II.45})$$

The joint distribution of $\{B_i : i = 0, \dots, n\}$ and its conditional distribution are defined by

$$\mathbb{P}\{B^n \in db^n\} \triangleq \mathbf{P}^P(db^n) = \int_{\mathbb{A}^n} \mathbf{P}^P(da^n, db^n), \quad n \in \mathbb{N}_0, \quad (\text{II.46})$$

$$\equiv \Pi_{0,n}^P(db^n) = \mu(db^{-1}) \otimes_{i=0}^n \Pi_i^P(db_i|b^{i-1}) \quad (\text{II.47})$$

$$\Pi_i^P(db_i|b^{i-1}) = \int_{\mathbb{A}^i} Q_i(db_i|b^{i-1}, a^i) \otimes P_i(da_i|a^{i-1}, b^{i-1}) \otimes \mathbf{P}^P(da^{i-1}|b^{i-1}), \quad i = 0, \dots, n. \quad (\text{II.48})$$

The above distributions are parametrized by either a fixed $B^{-1} = b^{-1} \in \mathbb{B}^{-1}$ or a fixed distribution $\mathbf{P}(db^{-1}) = \mu(db^{-1})$.

⁷The superscript notation, i.e., $\mathbf{P}^P, \mathbf{E}^P$ is used to track the dependence of the distribution and expectation on the channel input distribution $P_i(da_i|a^{i-1}, b^{i-1})$, $i = 0, \dots, n$.

FTFI Capacity. Directed information (pay-off) $I(A^n \rightarrow B^n)$ is defined by

$$I(A^n \rightarrow B^n) \triangleq \sum_{i=0}^n \mathbf{E}^P \left\{ \log \left(\frac{dQ_i(\cdot|B^{i-1}, A^i)}{d\Pi_i^P(\cdot|B^{i-1})}(B_i) \right) \right\} \quad (\text{II.49})$$

$$= \sum_{i=0}^n \int_{\mathbb{A}^i \times \mathcal{B}^i} \log \left(\frac{dQ_i(\cdot|b^{i-1}, a^i)}{d\Pi_i^P(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^P(da^i, db^i) \quad (\text{II.50})$$

$$\equiv \mathbb{I}_{A^n \rightarrow B^n}(P_i, Q_i, : i = 0, 1, \dots, n) \quad (\text{II.51})$$

where the notation (II.51) illustrates that $I(A^n \rightarrow B^n)$ is a functional of the two sequences of conditional distributions, $\{P_i(da_i|a^{i-1}, b^{i-1}), Q_i(db_i|b^{i-1}, a^i) : i = 0, 1, \dots, n\}$, and the initial distribution, which uniquely define the joint distribution, the marginal and conditional distributions $\{\mathbf{P}(da^i, db^i), \Pi_{0,i}^P(db^i), \Pi_i^P(db_i|b^{i-1}) : i = 0, 1, \dots, n\}$.

Clearly, (II.50) includes formulations with respect to probability density functions and probability mass functions.

Transmission Cost. The cost of transmitting and receiving symbols $a^n \in \mathbb{A}^n, b^n \in \mathbb{B}^n$ over the channel is a measurable function $c_{0,n} : \mathbb{A}^n \times \mathbb{B}^{n-1} \mapsto [0, \infty)$. The set of channel input distributions with transmission cost is defined by

$$\begin{aligned} \mathcal{P}_{[0,n]}(\kappa) \triangleq & \left\{ P_i(da_i|a^{i-1}, b^{i-1}) \in \mathcal{K}(\mathbb{A}_i|\mathbb{A}^{i-1} \times \mathbb{B}^{i-1}), i = 0, \dots, n : \right. \\ & \left. \frac{1}{n+1} \mathbf{E}^P(c_{0,n}(A^n, B^{n-1})) \leq \kappa \right\} \subset \mathcal{P}_{[0,n]}, \quad c_{0,n}(a^n, b^{n-1}) \triangleq \sum_{i=0}^n \gamma_i(T^i a^n, T^i b^{n-1}), \kappa \in [0, \infty) \end{aligned} \quad (\text{II.52})$$

where $\mathbf{E}^P\{\cdot\}$ denotes expectation with respect to the the joint distribution, and superscript “P” indicates its dependence on the choice of conditional distribution $\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, \dots, n\} \in \mathcal{P}_{[0,n]}$.

The characterization of feedback capacity $C_{A^n \rightarrow B^\infty}^{FB}(\kappa)$, is investigated as a consequence of the following definition of FTFI capacity characterization.

Definition II.1. (*Extremum problem with feedback*)

Given any channel distribution from the class $\mathcal{C}_{[0,n]}$, find the Information Structure of the optimal channel input distribution $\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, \dots, n\} \in \mathcal{P}_{[0,n]}(\kappa)$ (assuming it exists) of the extremum problem defined by

$$C_{A^n \rightarrow B^n}^{FB}(\kappa) \triangleq \sup_{\{P_i(da_i|a^{i-1}, b^{i-1}) : i=0, \dots, n\} \in \mathcal{P}_{[0,n]}(\kappa)} I(A^n \rightarrow B^n), \quad I(A^n \rightarrow B^n) = (\text{II.50}). \quad (\text{II.53})$$

If no transmission cost is imposed the optimization in (II.53) is carried out over $\mathcal{P}_{[0,n]}$, and $C_{A^n \rightarrow B^n}^{FB}(\kappa)$ is replaced by $C_{A^n \rightarrow B^n}^{FB}$.

Clearly, for each time i the largest information structure of the channel input conditional distribution of extremum problem $C_{A^n \rightarrow B^n}^{FB}(\kappa)$ is $\mathcal{S}_i^P \triangleq \{a^{i-1}, b^{i-1}\}, i = 1, \dots, n, \mathcal{S}_0^P \triangleq \{b^{-1}\}$.

Variational Equality of Directed Information. Often, in extremum problems of information theory, upper or lower bounds are introduced and then shown to be achievable over specific sets of distributions, such as, in entropy maximization with and without constraints, etc. In any extremum problem of capacity with feedback (resp. without feedback), identifying achievable upper bounds on directed information $I(A^n \rightarrow B^n)$ (resp. mutual information $I(A^n; B^n)$) is not an easy task. However, by invoking a variational equality of directed information [38] (resp. mutual information [39]), such achievable upper bounds can be identified.

Indeed, Step 2 of the proposed Two Step procedure (discussed in Section I) is based on utilizing the variation equality of directed information, given in the next theorem.

Theorem II.1. (*Variational Equality-Theorem IV.1 in [38].*)

Given a channel input conditional distribution $\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, 1, \dots, n\} \in \mathcal{P}_{[0,n]}$, a channel distribution $\{Q_i(db_i|b^{i-1}, a^i) : i = 0, 1, \dots, n\} \in \mathcal{C}_{[0,n]}$, and the initial distribution $\mu(db^{i-1})$, define the corresponding joint and marginal distributions by (II.44)-(II.48).

Let $\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i = 0, \dots, n\}$ be any arbitrary distribution. Then the following variational equality holds.

$$I(A^n \rightarrow B^n) = \inf_{\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i=0,1,\dots,n\}} \sum_{i=0}^n \int_{\mathbb{A}^i \times \mathbb{B}^i} \log \left(\frac{dQ_i(\cdot|b^{i-1}, a^i)}{dV_i(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^P(da^i, db^i) \quad (\text{II.54})$$

and the infimum in (II.54) is achieved at $V_i(db_i|b^{i-1}) = \Pi_i^P(db_i|b^{i-1}), i = 0, \dots, n$ given by (II.46)-(II.48).

The implications of variational equality (II.54) are illustrated via the following identity.

For any arbitrary distribution $\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i = 0, \dots, n\}$, the following identities hold.

$$\sum_{i=0}^n \mathbf{E}^P \left\{ \log \left(\frac{dQ_i(\cdot|B^{i-1}, A^i)}{dV_i(\cdot|B^{i-1})}(B_i) \right) \right\} = \sum_{i=0}^n \mathbf{E}^P \left\{ \log \left(\frac{dQ_i(\cdot|B^{i-1}, A^i)}{d\Pi_i^P(\cdot|B^{i-1})}(B_i) \right) \right\} + \sum_{i=0}^n \int_{\mathbb{B}^i} \log \left(\frac{d\Pi_i^P(\cdot|b^{i-1})}{V_i(\cdot|b^{i-1})}(b_i) \right) \Pi_{0,i}^P(db^i) \quad (\text{II.55})$$

$$= I(A^n \rightarrow B^n) + \sum_{i=0}^n \int_{\mathbb{B}^i} \log \left(\frac{d\Pi_i^P(\cdot|b^{i-1})}{V_i(\cdot|b^{i-1})}(b_i) \right) \Pi_{0,i}^P(db^i). \quad (\text{II.56})$$

Note that the second right hand side term in (II.56) is the sum of relative entropy terms between the marginal distribution $\Pi_i^P(db_i|b^{i-1})$ defined by the joint distribution $\mathbf{P}^P(da^i, db^i)$ (i.e., the correct conditional channel output distribution) and any arbitrary distribution $V_i(db_i|b^{i-1})$ (i.e., incorrect channel output conditional distribution) for $i = 0, 1, \dots, n$.

Identity (II.56) implies the minimization of its left hand side over any arbitrary channel output distribution $\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i = 0, \dots, n\}$ occurs at $V_i(db_i|b^{i-1}) = \Pi_i^P(db_i|b^{i-1}), i = 0, \dots, n$, i.e., when the relative entropy terms are zero, giving (II.54).

The point to be made regarding the above variational equality is that the characterization of the FTFI capacity can be transformed, for any arbitrary distribution $\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i = 0, \dots, n\}$, to the sequential equivalent $\sup \inf \{\cdot\}$ problem

$$C_{A^n \rightarrow B^n}^{FB}(\kappa) = \sup_{\mathcal{P}_{[0,n]}(\kappa)} \inf_{\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i=0,1,\dots,n\}} \sum_{i=0}^n \mathbf{E}^P \left\{ \log \left(\frac{dQ_i(\cdot|B^{i-1}, A^i)}{dV_i(\cdot|B^{i-1})}(B_i) \right) \right\} \quad (\text{II.57})$$

Then by removing the infimum in (II.57) an upper bound is identified, which together with stochastic optimal control techniques, is shown to be achievable over specific subsets of the set of all channel input conditional distributions satisfying conditional independence $\{\mathbf{P}(da_i|\mathcal{I}_i^P), \mathcal{I}_i^P \subseteq \{a^{i-1}, b^{i-1}\} : i = 0, \dots, n\} \subseteq \{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, \dots, n\}$ and the average transmission cost constraint. In fact, the characterizations of the FTFI capacity formulas for various channels and transmission cost functions discussed in this paper utilize this observation.

III. CHARACTERIZATION OF FTFI CAPACITY

The Two-Step Procedure. The identification of the information structures of the optimal channel input conditional distributions and the corresponding characterizations of the FTFI capacity $C_{A^n \rightarrow B^n}^{FB}$ and feedback capacity $C_{A^n \rightarrow B^n}^{FB}(\kappa)$, are determined by applying the following steps.

Step 1 Apply stochastic optimal control techniques with relaxed or randomized strategies (conditional distributions) [32], [40], [41], to show a certain joint process which generates the information structure of the channel input conditional distribution is an extended Markov process. This step implies the optimal channel input distribution occurs in specific subsets $\overline{\mathcal{P}}_{[0,n]} \subset \mathcal{P}_{[0,n]}$ or $\overline{\mathcal{P}}_{[0,n]}(\kappa) \subset \mathcal{P}_{[0,n]}(\kappa)$, which satisfy conditional independence.

Step 2 Apply variational equality of directed information given in Theorem II.1 ([38], Theorem I.V.1), to pay-off $I(A^n \rightarrow B^n)$, together with stochastic optimal control techniques, to identify upper bounds which are achievable over specific subsets $\overset{\circ}{\mathcal{P}}_{[0,n]} \subset \overline{\mathcal{P}}_{[0,n]}$ or $\overset{\circ}{\mathcal{P}}_{[0,n]}(\kappa) \subset \overline{\mathcal{P}}_{[0,n]}(\kappa)$, which satisfy a further conditional independence.

For certain channel distributions and instantaneous transmission cost functions, Step 1 is sufficient to identify the information structures of channel input distributions (i.e., Class A channels and transmission cost functions), and to characterize the FTFI capacity, while for others, Step 1 may serve as an intermediate step prior to applying Step 2. For example, if the channel

distribution is of limited memory with respect to channel outputs, i.e., of the Class B, by applying Step 2 an upper bound on the FTFI capacity is obtained, which together with stochastic optimal control techniques, it is shown to be achievable over channel input distributions with limited memory on channel outputs.

It is also possible to apply Steps 1 and 2 jointly; this will be illustrated in specific applications.

Step 1 is a generalization of equivalent methods often applied in stochastic optimal Markov decision or control problems to show that optimizing a pay-off [33], [42] over all possible non-Markov policies or strategies, occurs in the smaller set of Markov policies. However, certain issues should be treated with caution, when stochastic optimal control techniques are applied in extremum problems of information theory. These are summarized in the next remark.

Comments on stochastic optimal control in relation to extremum problems of information theory. In fully observable stochastic optimal control theory [32], one is given a controlled process $\{X_i : i = 0, \dots, n\}$, often called the state process taking values in $\{\mathbb{X}_i : i = 0, \dots, n\}$, affected by a control process $\{U_i : i = 0, \dots, n\}$ taking values in $\{\mathbb{U}_i : i = 0, \dots, n\}$, and the corresponding control object $\mathcal{P}_{[0,n]}^{CO} \triangleq \{\mathbf{P}_{U_i|U^{i-1}, X^i} : i = 0, \dots, n\}$ and the controlled object $\mathcal{C}_{[0,n]}^{CO} \triangleq \{\mathbf{P}_{X_i|X^{i-1}, U^{i-1}} : i = 0, \dots, n\}$. Often, the controlled object is Markov conditional on the past control values, that is, $\mathbf{P}_{X_i|X^{i-1}, U^{i-1}} = \mathbf{P}_{X_i|X_{i-1}, U_{i-1}} - a.a.(x^{i-1}, u^{i-1})$, $i = 0, \dots, n$. Such Markov controlled objects are often induced by discrete recursions

$$X_{i+1} = f_i(X_i, U_i, V_i), \quad X_0 = x_0, \quad i = 0, \dots, n, \quad (\text{III.58})$$

where $\{V_i : i = 0, \dots, n\}$ is an independent noise process taking values in $\{\mathbb{V}_i : i = 0, \dots, n\}$, independent of the initial state X_0 . Denote the set of such Markov distributions or controlled objects by $\mathcal{C}_{[0,n]}^{CO-M} \triangleq \{\mathbf{P}_{X_i|X_{i-1}, U_{i-1}} : i = 0, \dots, n\}$.

In stochastic optimal control theory, one is given a sample pay-off function, often of additive form, defined by

$$l : \mathbb{X}^n \times \mathbb{U}^n \mapsto (-\infty, \infty], \quad l(x^n, u^n) \triangleq \sum_{i=0}^n \ell(u_i, x_i) \quad (\text{III.59})$$

where the functions $\{\ell_i(\cdot, \cdot) : i = 0, \dots, n\}$ are fixed and independent of the control object $\{\mathbf{P}_{U_i|U^{i-1}, X^i} : i = 0, \dots, n\}$.

Given the Markov distribution $\mathbf{P}_{X_i|X_{i-1}, U_{i-1}} : i = 0, \dots, n$, the objective is to optimize the average of the sample path pay-off over all non-Markov strategies in $\mathcal{P}_{[0,n]}^{CO}$, i.e.,

$$J_{0,n}^F(\mathbf{P}_{U_i|U^{i-1}, X^i}^*, i = 0, \dots, n) \triangleq \inf_{\mathcal{P}_{[0,n]}^{CO}} \mathbf{E} \left\{ \sum_{i=0}^n \ell(U_i, X_i) \right\}. \quad (\text{III.60})$$

Two features of stochastic optimal control which are distinct from any extremum problem of directed information are discussed below.

Feature 1. Stochastic optimal control formulations pre-suppose, that additional state variables are introduced, prior to arriving at the Markov controlled object $\{\mathbf{P}_{X_i|X_{i-1}, U_{i-1}} : i = 0, \dots, n\}$ or the discrete recursion (III.58), and the pay-off (III.59). Specifically, the final formulation (III.60), pre-supposes the Markov controlled object $\{\mathbf{P}_{X_i|X_{i-1}, U_{i-1}} : i = 0, \dots, n\}$ is obtained as follows. The state variables which constitute the complete state process $\{X_i : i = 0, \dots, n\}$ may be due to a noise process which was not independent and converted into an independent noise process via state augmentation, and/or any dependence on past information, and converted to a Markov controlled object via state augmentation, and due to a non-single letter dependence, for each i , of the the sample pay-off functions $\ell_i(\cdot, \cdot)$, which was converted into single letter dependence, i.e. (x_i, u_i) , by additional state augmentation, so that the controlled object is Markov. Such examples are given in [43] for deterministic or non-randomized strategies, defined by

$$\mathcal{C}_{[0,n]}^{CO} \triangleq \{e_i : \mathbb{U}^{i-1} \times \mathbb{X}^i \mapsto \mathbb{U}_i, \quad i = 0, \dots, n : \quad u_i = e_i(u^{i-1}, x^i), i = 0, \dots, n\}. \quad (\text{III.61})$$

In view of the Markovian property of the controlled object, i.e., given by $\mathbf{P}_{X_i|X_{i-1}, U_{i-1}}, i = 0, \dots, n$, then the optimization in

(III.60) reduces to the following optimization problem over Markov strategies.

$$J_{0,n}^F(\mathbf{P}_{U_i|U^{i-1},X^i}^*, i=0, \dots, n) = J_{0,n}^M(\mathbf{P}_{U_i|X_i}^*, i=0, \dots, n) \triangleq \inf_{\mathbf{P}_{U_i|X_i}, i=0, \dots, n} \mathbf{E} \left\{ \sum_{i=0}^n \ell(U_i, X_i) \right\}. \quad (\text{III.62})$$

This further implies that the control process $\{X_i : i=0, \dots, n\}$ is Markov, i.e., it satisfies $\mathbf{P}_{X_i|X^{i-1}} = \mathbf{P}_{X_i|X_{i-1}}, i=0, \dots, n$. On the other hand, if $\mathbf{P}_{X_i|X^{i-1}} = \mathbf{P}_{X_i|X_{i-1}}, i=0, \dots, n$ then (III.62) holds.

Feature 2. Given a general controlled object $\{\mathbf{P}_{X_i|X^{i-1}, U^{i-1}} : i=0, \dots, n\}$ non necessarily Markov, one of the fundamental results of classical stochastic optimal control is that optimizing the pay-off $\mathbf{E} \left\{ \sum_{i=0}^n \ell(U_i, X_i) \right\}$ over randomized strategies $\mathcal{P}_{[0,n]}^{CO}$ does not incur a better performance than optimizing it over non-randomized strategies $\mathcal{E}_{[0,n]}^{CO}$, i.e.,

$$J_{0,n}^F(\mathbf{P}_{U_i|U^{i-1}, X^i}^*, i=0, \dots, n) = \inf_{\mathcal{E}_{[0,n]}^{CO}} \mathbf{E} \left\{ \sum_{i=0}^n \ell(U_i, X_i) \right\} \quad (\text{III.63})$$

$$= \inf_{g_i(X_i): i=0, \dots, n} \mathbf{E}^g \left\{ \sum_{i=0}^n \ell(U_i, X_i) \right\} \quad \text{if } \mathbf{P}_{X_i|X^{i-1}, U^{i-1}} = \mathbf{P}_{X_i|X_{i-1}, U_{i-1}} - a.a., i=0, \dots, n. \quad (\text{III.64})$$

Step 2, i.e., the application of variational equality, discussed in Section I, is specific to information theoretic pay-off functionals and does not have a counterpart to any of the common pay-off functionals of stochastic optimal control problems [33], [42]. This is due to the fact that, unlike stochastic optimal control problems (discussed above, feature (1)), any extremum problem of feedback capacity involves directed information density $\iota_{A^n \rightarrow B^n}(a^n, b^n) \equiv \iota_{A^n \rightarrow B^n}^P(a^n, b^n)$ defined by (I.7), which is not a fixed functional, but depends on the channel output transition probability distribution $\{\mathbf{P}_{B_i|B^{i-1}} \equiv \Pi_i^P(db_i|b^{i-1}) : i=0, \dots, n\}$ defined by (II.48), which depends on the channel distribution, and the channel input distribution chosen to maximize directed information $I(A^n \rightarrow B^n)$. The nonlinear dependence of the directed information density makes extremum problems of directed information, distinct from extremum problems of classical stochastic optimal control.

This implies step 2 or more specifically, the variational equalities of directed information and mutual information, are key features of information theoretic pay-off functionals. Often, these variational equalities need to be incorporated into any extremum problems of deriving achievable bounds, such as, in extremum problems of feedback capacity and capacity without feedback, much as, it is often done when deriving achievable bounds, based on the entropy maximizing properties of distributions (i.e., Gaussian distributions).

Feature (2), i.e., (III.63) and (III.64), do not have counters part in any extremum problem of directed information or mutual information, that is, optimizing directed information over channel input distributions is not equivalent to optimizing directed information over deterministic non-randomized strategies. In fact, by definition the sequence of codes defined by (I.3) are randomized strategies, and if these are specialized to non-randomized strategies, i.e., by removing their dependence on the randomly generated messages, $W \in \mathcal{M}_n$, then directed information is zero, i.e., $I(A^n \rightarrow B^n) = 0$ if $a_i = e_i(a^{i-1}, b^{i-1}), i=0, \dots, n$.

A. Channels Class A and Transmission Costs Class A or B

First, the preliminary steps of the derivation of the characterization of FTFI capacity for any channel distributions of Class A, (I.15), without transmission cost are introduced, to gain insight into the derivations of information structures, without the need to introduce excessive notation. The analogies and differences between stochastic optimal control theory and extremum problems of directed information, as discussed above, are made explicit, throughout the derivations, when tools from stochastic optimal control are applied to the directed information density pay-off.

Introduce the following definition of channel input distributions satisfying conditional independence.

Definition III.1. (Channel input distributions for Class A channels and transmission cost functions)

Define the restricted class of channel input distributions $\overline{\mathcal{P}}_{[0,n]}^A \subset \mathcal{P}_{[0,n]}$ satisfying conditional independence by

$$\begin{aligned} \overline{\mathcal{P}}_{[0,n]}^A &\triangleq \left\{ \{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, 1, \dots, n\} \in \mathcal{P}_{[0,n]} : \right. \\ &\quad \left. P_i(da_i|a^{i-1}, b^{i-1}) = \mathbf{P}_{A_i|B^{i-1}}(da_i|b^{i-1}) \equiv \pi_i(da_i|b^{i-1}) - a.a.(a^{i-1}, b^{i-1}), i = 0, 1, \dots, n \right\}. \end{aligned} \quad (\text{III.65})$$

Similarly, for transmission cost functions $\gamma_i(T^i a^n, T^i b^{n-1}) = \gamma_i^A(a_i, b^{i-1})$ or $\gamma_i(T^i a^n, T^i b^{n-1}) = \gamma_i^{B,K}(a_i, b_{i-K}^{i-1}), i = 0, \dots, n$, define

$$\overline{\mathcal{P}}_{[0,n]}^A(\kappa) \triangleq \overline{\mathcal{P}}_{[0,n]}^A \cap \mathcal{P}_{[0,n]}(\kappa). \quad (\text{III.66})$$

From the definition of directed information $I(A^n \rightarrow B^n)$ given by (II.50), and utilizing the channel distribution (I.15), the FTFI capacity is defined by

$$C_{A^n \rightarrow B^n}^{FB,A} \triangleq \sup_{\{P_i(da_i|a^{i-1}, b^{i-1}) : i=0, \dots, n\} \in \overline{\mathcal{P}}_{[0,n]}^A} \sum_{i=0}^n \mathbf{E}^P \left\{ \log \left(\frac{dQ_i(\cdot|B^{i-1}, A_i)}{d\Pi_i^P(\cdot|B^{i-1})}(B_i) \right) \right\} \quad (\text{III.67})$$

where the channel output transition probability defined by (II.48), is given by the following expressions.

$$\Pi_i^P(db_i|b^{i-1}) = \int_{\mathbb{A}^i} Q_i(db_i|b^{i-1}, a^i) \otimes \mathbf{P}^P(da^i|b^{i-1}), \quad i = 0, \dots, n \quad (\text{III.68})$$

$$= \int_{\mathbb{A}^i} Q_i(db_i|b^{i-1}, a_i) \otimes P_i(da_i|a^{i-1}, b^{i-1}) \otimes \mathbf{P}^P(da^{i-1}|b^{i-1}) \quad (\text{III.69})$$

$$\stackrel{(\alpha)}{=} \Pi_i^{\pi_i}(db_i|b^{i-1}) \triangleq \int_{\mathbb{A}^i} Q_i(db_i|b^{i-1}, a_i) \otimes \pi_i(da_i|b^{i-1}) \quad (\text{III.70})$$

$$\text{if } P_i(da_i|a^{i-1}, b^{i-1}) = \mathbf{P}(da_i|b^{i-1}) \equiv \pi_i(da_i|b^{i-1}) - a.a.(a^{i-1}, b^{i-1}) \quad i = 0, \dots, n \quad (\text{III.71})$$

Note that identity (α) holds if it can be shown that conditional independence (III.71) holds for any candidate $\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, 1, \dots, n\} \in \mathcal{P}_{[0,n]}$ maximizing $I(A^n \rightarrow B^n)$; the superscript notation, $\Pi_i^{\pi_i}(db_i|b^{i-1})$, indicates the dependence of $\Pi_i^P(db_i|b^{i-1})$ on conditional distribution $\mathbf{P}(da_i|b^{i-1}) \equiv \pi_i(da_i|b^{i-1})$ (instead on $\{P_j(da_j|a^{j-1}, b^{j-1}) : j = 0, 1, \dots, i\}$), for $i = 1, \dots, n$.

It is important to note that one cannot assume $\Pi_i^P(db_i|b^{i-1}) = \int_{\mathbb{A}^i} Q_i(db_i|b^{i-1}, a_i) \otimes \pi_i(da_i|b^{i-1}) \equiv \Pi_i^{\pi_i}(db_i|b^{i-1})$, that is, (III.69) is given by (III.70) without proving that such restriction of conditional distributions is a property of the channel input distribution which maximizes directed information $I(A^n \rightarrow B^n)$, because the marginal distribution $\Pi^P(db^n)$ is uniquely defined from the joint distribution $\mathbf{P}^P(da^n, db^n) = \otimes_{i=0}^n \left(Q_i(db_i|b^{i-1}, a_i) \otimes P_i(da_i|a^{i-1}, b^{i-1}) \right)$. The derivation of Theorem 1 in [29] and Theorem 1 in [22] for the problems considered by the authors, should be read with caution, to account for the above feature, in order to show the supremum over all channel input conditional distributions occurs in the smaller set, satisfying a conditional independence condition, which is analogous to (III.71).

Suppose (III.71) holds (its validity is shown in Theorem III.1). Then the expectation $\mathbf{E}^P\{\cdot\}$ in (III.67) with respect to the joint distribution simplifies as follows.

$$\mathbf{P}^P(da_i, db^i) = \mathbf{P}^P(da_i, db^i) = Q_i(db_i|b^{i-1}, a_i) \otimes \mathbf{P}(da_i|b^{i-1}) \otimes \mathbf{P}^P(db^{i-1}), \quad i = 0, \dots, n \quad (\text{III.72})$$

$$= \mathbf{P}^{\pi_i}(da_i, db^i) \quad \text{if (III.71) holds, } i = 0, 1, \dots, n, \quad (\text{III.73})$$

$$= Q_i(db_i|b^{i-1}, a_i) \otimes \pi_i(da_i|b^{i-1}) \otimes \Pi_{0,i-1}^{\pi_0, \dots, \pi_{i-1}}(db^{i-1}), \quad i = 0, \dots, n, \quad (\text{III.74})$$

$$\Pi_{0,i-1}^{\pi_0, \dots, \pi_{i-1}}(db^{i-1}) = \Pi_{i-1}^{\pi_{i-1}}(db_{i-1}|b^{i-2}) \otimes \Pi_{0,i-2}^{\pi_0, \dots, \pi_{i-2}}(db^{i-2}), \quad i = 0, \dots, n \quad (\text{III.75})$$

where the superscript notation, $\mathbf{P}^{\pi_i}(da_i, b^i)$, indicates the dependence of joint distribution $\mathbf{P}^P(da_i, b^i)$ on $\{\pi_j(da_j|b^{j-1}) : j = 0, 1, \dots, i\}$, for $i = 0, \dots, n$. Clearly, if (III.70) holds, for each i , the controlled conditional distribution-controlled object, $\Pi_i^P(db_i|b^{i-1}) = \Pi_i^{\pi_i}(db_i|b^{i-1})$, depends on the channel distribution $Q_i(db_i|b^{i-1}, a_i)$ and the control conditional distribution-control object, $\pi_i(da_i|b^{i-1})$, for $i = 0, 1, \dots, n$.

Thus, the following holds.

- **Channel Class A.1, (I.15):** If the maximizing channel input conditional distribution satisfies conditional independence $P_i(da_i|a^{i-1}, b^{i-1}) = \pi_i(da_i|b^{i-1}) - a.a.(a^{i-1}, b^{i-1}), i = 0, \dots, n$, then directed information (II.50) is a functional of $\{Q_i(db_i|b^{i-1}, a_i), \pi_i(da_i|b^{i-1}) : i = 0, \dots, n\}$ and it is given by

$$I(A^n \rightarrow B^n) = \sum_{i=0}^n I(A_i; B_i | B^{i-1}) = \sum_{i=0}^n \mathbf{E}^\pi \left\{ \log \left(\frac{dQ_i(\cdot | B^{i-1}, A_i)}{d\pi_i(\cdot | B^{i-1})}(B_i) \right) \right\} \quad (\text{III.76})$$

$$= \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot | b^{i-1}, a_i)}{d\pi_i(\cdot | b^{i-1})}(b_i) \right) Q_i(db_i | b^{i-1}, a_i) \otimes \pi_i(da_i | b^{i-1}) \otimes \Pi_{0,i-1}^{\pi_0, \dots, \pi_{i-1}}(db^{i-1}) \quad (\text{III.77})$$

$$\equiv \mathbb{I}_{A^n \rightarrow B^n}(\pi_i, Q_i : i = 0, \dots, n) \quad (\text{III.78})$$

where $\mathbf{E}^\pi\{\cdot\}$ indicates that the joint distribution over which expectation is taken depends on the sequence of conditional distributions $\{\pi_j(da_j | b^{j-1}) : j = 0, 1, \dots, i\}$, for $i = 0, \dots, n$.

By (III.76), since the expectation is taken with respect to joint distribution (III.74), the distribution $\{\pi_i(da_i | b^{i-1}) : i = 0, \dots, n\}$ is indeed the control object (conditional distribution), chosen to control the conditional distribution of the channel output process, $\{\Pi_i^{\pi_i}(db_i | b^{i-1}) : i = 0, 1, \dots, n\}$. By analogy with stochastic optimal control with randomized strategies, for each i , the conditional distribution $\Pi_i^{\pi_i}(db_i | b^{i-1})$ is affected by the control object $\pi_i(da_i | b^{i-1})$, for $i = 0, \dots, n$, and this is chosen to influence the pay-off (III.78), which is a functional of $\{\pi_i(da_i | b^{i-1}) : i = 0, \dots, n\}$ (since the channel is fixed).

Next, it is shown that (III.70) is indeed valid, i.e., the maximization of $I(A^n \rightarrow B^n)$ over $\{P_i(da_i | a^{i-1}, b^{i-1}) : i = 0, \dots, n\}$ occurs in the smaller set $\overline{\mathcal{P}}_{[0,n]}^A \subset \mathcal{P}_{[0,n]}$.

Theorem III.1. (Characterization of FTFI capacity for channels of class A)

Suppose the channel distribution is of Class A defined by (I.15). Then the following hold.

Part A. The maximization of $I(A^n \rightarrow B^n)$ over $\mathcal{P}_{[0,n]}$ occurs in $\overline{\mathcal{P}}_{[0,n]}^A \subset \mathcal{P}_{[0,n]}$ and the characterization of FTFI capacity is given by the following expression.

$$C_{A^n \rightarrow B^n}^{FBA} = \sup_{\{\pi_i(da_i | b^{i-1}) \in \mathcal{M}(\mathbb{A}_i) : i=0, \dots, n\}} \sum_{i=0}^n \mathbf{E}^\pi \left\{ \log \left(\frac{dQ_i(\cdot | B^{i-1}, A_i)}{d\pi_i(\cdot | B^{i-1})}(B_i) \right) \right\} \quad (\text{III.79})$$

where $\Pi_i^{\pi_i}(db_i | b^{i-1}) = \int_{\mathbb{A}_i} Q_i(db_i | b^{i-1}, a_i) \otimes \pi_i(da_i | b^{i-1})$ and the joint distribution over which $\mathbf{E}^\pi\{\cdot\}$ is taken is $\{\mathbf{P}^\pi(da_i, db^i) : i = 0, \dots, n\}$ defined by (III.74).

Part B. Suppose the following two conditions hold.

$$(a) \quad \gamma_i(T^i a^n, T^i b^{n-1}) = \gamma_i^A(a_i, b^{i-1}) \quad \text{or} \quad \gamma_i(T^i a^n, T^i b^{n-1}) = \gamma_i^{B,K}(a_i, b_{i-K}^{i-1}), \quad i = 0, \dots, n, \quad (\text{III.80})$$

$$(b) \quad C_{A^n \rightarrow B^n}^{FBA}(\kappa) \triangleq \sup_{\{P_i(da_i | a^{i-1}, b^{i-1}) : i=0, \dots, n\} \in \mathcal{P}_{[0,n]}(\kappa)} I(A^n \rightarrow B^n) \quad (\text{III.81})$$

$$= \inf_{s \geq 0} \sup_{\{P_i(da_i | a^{i-1}, b^{i-1}) : i=0, \dots, n\} \in \mathcal{P}_{[0,n]}(\kappa)} \left\{ I(A^n \rightarrow B^n) - s \left\{ \mathbf{E}^P \left(c_{0,n}(A^n, B^{n-1}) \right) - \kappa(n+1) \right\} \right\}. \quad (\text{III.82})$$

The maximization of $I(A^n \rightarrow B^n)$ over channel input distributions with transmission cost $\{P_i(da_i | a^{i-1}, b^{i-1}) : i = 1, \dots, n\} \in \mathcal{P}_{[0,n]}(\kappa)$ occurs in $\overline{\mathcal{P}}_{[0,n]}^A(\kappa)$, and the FTFI capacity is given by the following expression.

$$C_{A^n \rightarrow B^n}^{FBA}(\kappa) = \sup_{\pi_i(da_i | b^{i-1}) \in \mathcal{M}(\mathbb{A}_i), i=0, \dots, n; \frac{1}{n+1} \mathbf{E}^\pi \left\{ c_{0,n}(A^n, B^{n-1}) \right\} \leq \kappa} \sum_{i=0}^n \mathbf{E}^\pi \left\{ \log \left(\frac{dQ_i(\cdot | B^{i-1}, A_i)}{d\pi_i(\cdot | B^{i-1})}(B_i) \right) \right\}. \quad (\text{III.83})$$

Proof: The derivation is based on expressing directed information as a functional of the channel input conditional distribution, identifying the explicit dependence of the sample path pay-off on appropriate state variable, and then showing the controlled object, which is defined using the state variable is Markov, i.e., as discussed earlier.

Part A. By the channel distribution assumption (I.15), the following equalities are obtained.

$$I(A^n \rightarrow B^n) = \sum_{i=0}^n \mathbf{E}^P \left\{ \log \left(\frac{dQ_i(\cdot|B^{i-1}, A^i)}{d\Pi_i^P(\cdot|B^{i-1})}(B_i) \right) \right\} \quad (\text{III.84})$$

$$\stackrel{(\alpha)}{=} \sum_{i=0}^n \int_{\mathbb{A}_i \times \mathbb{B}^i} \log \left(\frac{dQ_i(\cdot|B^{i-1}, A_i)}{d\Pi_i^P(\cdot|B^{i-1})}(B_i) \right) \mathbf{P}^P(dA_i, dB^i) \quad (\text{III.85})$$

$$\stackrel{(\beta)}{=} \sum_{i=0}^n \mathbf{E}^P \left\{ \log \left(\frac{dQ_i(\cdot|B^{i-1}, A_i)}{d\Pi_i^P(\cdot|B^{i-1})}(B_i) \right) \right\} \quad (\text{III.86})$$

$$\stackrel{(\gamma)}{=} \sum_{i=1}^n \mathbf{E}^P \left\{ \mathbf{E}^P \left\{ \log \left(\frac{dQ_i(\cdot|B^{i-1}, A_i)}{d\Pi_i^P(\cdot|B^{i-1})}(B_i) \right) \middle| A^i, B^{i-1} \right\} \right\} \quad (\text{III.87})$$

$$\stackrel{(\delta)}{=} \sum_{i=0}^n \mathbf{E}^P \left\{ \mathbf{E}^P \left\{ \log \left(\frac{dQ_i(\cdot|B^{i-1}, A_i)}{\Pi_i^P(\cdot|B^{i-1})}(B_i) \right) \middle| A_i, B^{i-1} \right\} \right\} \quad (\text{III.88})$$

$$\stackrel{(\epsilon)}{=} \sum_{i=0}^n \mathbf{E}^P \left\{ \ell_i^P(A_i, S_i) \right\}, \quad S_j \triangleq B^{j-1}, \quad j = 0, \dots, n, \quad (\text{III.89})$$

$$\ell_j^P(a_j, s_j) \triangleq \int_{\mathbb{B}_j} \log \left(\frac{dQ_j(\cdot|s_j, a_j)}{d\Pi_j^P(\cdot|s_j)}(b_j) \right) Q_j(db_j|s_j, a_j), \quad j = 0, \dots, n \quad (\text{III.90})$$

where

(α) is due to the channel distribution assumption (I.15);

(β) is by definition;

(γ) is due to a property of expectation;

(δ) is due to the channel distribution (I.15);

(ϵ) is by definition of conditional expectation for the measurable function $\ell_i^P(\cdot, \cdot)$ defined by (III.90).

The validity of the claim that the optimal channel input conditional distribution belongs to the class $\overline{\mathcal{D}}_{[0,n]}^A$, establishing validity of the claimed identity (III.70), and consequently validity of (III.76)-(III.78), is shown as follows. Since for each i , the channel $Q_i(db_i|b^{i-1}, a_i)$ and the pay-off function $\ell_i^P(a_i, s_i) \equiv \ell_i^P(a_i, b^{i-1})$ in (III.89) depend on $s_i \triangleq b^{i-1}$ for $i = 0, 1, \dots, n$, then $\{S_i \triangleq B^{i-1} : i = 0, \dots, n\}$ is the controlled process, control by the control process $\{A_i : i = 0, \dots, n\}$ (see discussion on stochastic optimal control). That is, $\{\mathbf{P}(ds_{i+1}|s^i, a^i) : i = 0, \dots, n-1\}$ is the controlled object. Next, we show the controlled process $\{S_i : i = 0, \dots, n\}$ is Markov, i.e., the transition probabilities $\{\mathbf{P}(ds_{i+1}|s^i) : i = 0, \dots, n-1\}$ are Markov, and the maximization of directed information occurs in the set $\overline{\mathcal{D}}_{[0,n]}^A$, defined by (III.65). By applying Bayes' theorem and using the definition of the channel distribution, the following conditional independence are easily shown.

$$\mathbf{P}(ds_{i+1}|s^i, a^i) = \mathbf{P}(ds_{i+1}|s_i, a_i), \quad i = 0, \dots, n-1, \quad (\text{III.91})$$

$$\mathbf{P}(ds_{i+1}|s^i) = \mathbf{P}(ds_{i+1}|s_i) = \int_{\mathbb{A}_i} \mathbf{P}(ds_{i+1}|s_i, a_i) \otimes \mathbf{P}(da_i|s_i). \quad (\text{III.92})$$

In fact, since the controlled object is Markov, i.e., (III.91) holds, the statement of the theorem follows directly from stochastic optimal control [32], [42], (see discussion on stochastic optimal control). Nevertheless, we give the complete derivation.

In view of the above identities the Markov process $\{S_i : i = 0, \dots, n\}$ satisfies the following identity.

$$\mathbf{P}^P(ds_{i+1}) = \int_{\mathbb{B}^{i-1} \times \mathbb{A}_i} \mathbf{P}(ds_{i+1}|s_i, a_i) \otimes \mathbf{P}(da_i|s_i) \otimes \mathbf{P}^P(ds_i) \implies \mathbf{P}^\pi(ds_{i+1}) = \int_{\mathbb{B}^{i-1} \times \mathbb{A}_i} \mathbf{P}(ds_{i+1}|s_i, a_i) \otimes \pi_i(da_i|s_i) \otimes \mathbf{P}^\pi(ds_i) \quad (\text{III.93})$$

where $\mathbf{P}^P(ds_{i+1}) = \mathbf{P}^{\pi_0, \dots, \pi_i}(ds_{i+1}) \equiv \mathbf{P}^\pi(ds_{i+1})$ follows by iterating the first equation in (III.93). Since, the process $\{S_i : i = 0, 1, \dots, n\}$ is Markov with transition probability given by the right hand side of (III.92), then for $i = 0, \dots, n-1$, the distribution $\mathbf{P}(ds_{i+1}|s_i)$ is controlled by the control object $\mathbf{P}(da_i|s_i) \equiv \mathbf{P}(da_i|b^{i-1})$. Clearly, (III.92) implies that any measurable function say, $\xi(s_i)$ of $s_i = b^{i-1}$ is affected by the control object $\mathbf{P}(da_i|s_i)$, and hence by (III.69), then $\{\xi_i(s_i) \triangleq \Pi_i^P(db_i|b^{i-1}) \equiv \Pi_i^{\pi_i}(db_i|b^{i-1}) : i = 0, \dots, n\}$, and this transition distribution is controlled by the control object $\{\pi_i(da_i|b^{i-1}) : i = 0, \dots, n\}$. Utilizing this in

(III.90) and (III.89), the following is obtained.

$$I(A^n \rightarrow B^n) = \sum_{i=0}^n \int_{\mathbb{A}_i \times \mathbb{B}^i} \log \left(\frac{dQ_i(\cdot|b^{i-1}, a_i)}{d\Pi_i^\pi(\cdot|b^{i-1})}(b_i) \right) Q_i(db_i|b^{i-1}, a_i) \otimes \pi(da_i|b^{i-1}) \otimes \mathbf{P}^\pi(db^{i-1}). \quad (\text{III.94})$$

Thus, the maximization in (III.94) over all channel input distributions is done by choosing the control object $\{\pi_i(da_i|b^{i-1}) : i = 0, \dots, n\}$ to control the conditional distribution $\{\Pi_i^{\pi_i}(db_i|b^{i-1}) : i = 1, \dots, n\}$, which for each i , depends on $\{b^{i-1}, \pi_i(da_i|b^{i-1})\}$, for $i = 0, \dots, n$. Hence, the maximizing object in (III.67) (if it exists), is of the form $P_i^*(da_i|a^{i-1}, b^{i-1}) = \pi_i^*(da_i|b^{i-1}) - a.a.(a^{i-1}, b^{i-1}), i = 1, \dots, n\}$, and (III.79) is obtained.

Part B. Since for each i , the channel conditional distribution $Q_i(db_i|\cdot, \cdot)$ is measurable with respect to $\mathcal{J}_i^Q = \{b^{i-1}, a_i\}$ and the transmission cost $\gamma_i(\cdot, \cdot) \equiv \gamma_i^A(\cdot, \cdot)$ or $\gamma_i(\cdot, \cdot) \equiv \gamma_i^{B,K}(\cdot, \cdot)$ is measurable with respect to $\mathcal{J}_i^Y = \{a_i, b^{i-1}\}$ or $\mathcal{J}_i^Y = \{a_i, b_{i-K}^{i-1}\}$ for $i = 0, 1, \dots, n$, the results follow directly from Part A, as follows. Consider the cost function $\{\gamma_i^A(a_i, b^{i-1}), i = 0, \dots, n\}$, and note that the average cost constraint can be expressed as follows.

$$\sum_{i=0}^n \mathbf{E}^P \left\{ \gamma_i^A(A_i, B^{i-1}) \right\} = \sum_{i=0}^n \mathbf{E}^P \left\{ \mathbf{E}^P \left\{ \gamma_i^A(A_i, B^{i-1}) \middle| A^i, B^{i-1} \right\} \right\} \quad (\text{III.95})$$

$$= \sum_{i=0}^n \mathbf{E}^\pi \left\{ \bar{\gamma}_i^{A,\pi}(S_i) \right\}, \quad \bar{\gamma}_j^{A,\pi}(s_j) \triangleq \int_{\mathbb{A}_j} \gamma_j^A(a_j, s_j) \pi_j(da_j|s_j), \quad j = 0, \dots, n. \quad (\text{III.96})$$

where the expectation $\mathbf{E}^\pi\{\cdot\}$ is taken with respect to $\mathbf{P}^\pi(ds_i)$ and this follows from Part A. Since the transmission cost constraint is expressed via (III.96) and depends on the distribution $\{\pi_i(da_i|s_i) : i = 0, \dots, n\}$ then the claim holds.

Alternatively, if condition (b) holds then the Lagrangian of the unconstraint problem (omitting the term $\kappa(n+1)$) is

$$I(A^n \rightarrow B^n) - s \mathbf{E}^P \left(c_{0,n}(A^n, B^{n-1}) \right) = \sum_{i=0}^n \mathbf{E}^P \left\{ \ell_i^P(A_i, S_i) - s \gamma_i^A(A_i, B^{n-1}) \right\} \quad (\text{III.97})$$

and the rest of the derivation follows from Part A. This completes the prove. For the cost function $\{\gamma_i^{B,K}(a_i, b_{i-K}^{i-1}), i = 0, \dots, n\}$, since b_{i-K}^{i-1} is the restriction of $s_i = b^{i-1}$ to a subsequence, the above conclusion holds as well. ■

Next, we give some comments regarding the previous theorem and discuss possible generalizations.

Remark III.1. (Some generalizations)

(1) Suppose the channel is $\{Q_i(db_i|b_{i-1}, a_i) : i = 0, \dots, n\}$ and the transmission cost function is $\gamma_i(T^i a^n, T^i b^{n-1}) = \gamma_i^A(a_i, b^{i-1}), i = 0, \dots, n$. Then the statements of Theorem III.1, Part B, remain valid with $Q_i(db_i|b^{i-1}, a_i)$ replaced by $Q_i(db_i|b_{i-1}, a_i)$ for $i = 0, \dots, n$, because the state is $s_i = b^{i-1}$, and this is determined from the dependence of the cost function $\gamma_i^A(a_i, b^{i-1})$ on s_i , for $i = 0, \dots, n$.

(2) Suppose in Theorem III.1, Part B, $\gamma_i(T^i a^n, T^i b^n) = \gamma_i^A(a_i, b^i), i = 0, \dots, n$, then from (III.95), (III.96) we have

$$\sum_{i=0}^n \mathbf{E}^P \left\{ \gamma_i^A(A_i, B^i) \right\} = \sum_{i=0}^n \mathbf{E}^\pi \left\{ \bar{\gamma}_i^{A,\pi}(S_i) \right\}, \quad \bar{\gamma}_j^{A,\pi}(s_j) \triangleq \int_{\mathbb{A}_j \times \mathbb{B}_j} \gamma_j^A(a_j, s_j) Q_j(db_j|s_j, a_j) \otimes \pi_j(da_j|s_j), \quad j = 0, \dots, n. \quad (\text{III.98})$$

Hence, the statements of Theorem III.1, Part B remain valid.

Remark III.2. (Equivalence of constraint and unconstraint problems)

The equivalence of constraint and unconstraint problems in Theorem III.1, follows from Lagrange's duality theory of optimizing convex functionals over convex sets [44]. Specifically, from [38], it follows that the set of distributions $\mathbf{P}^{C1}(da^n|b^{n-1}) \triangleq \otimes_{i=0}^n P_i(da_i|a^{i-1}, b^{i-1}) \in \mathcal{M}(\mathbb{A}^n)$ is convex, and this uniquely defines $\mathcal{P}_{[0,n]}$ and vice-versa, directed information as a functional of $\mathbf{P}^{C1}(da^n|b^{n-1}) \in \mathcal{M}(\mathbb{A}^n)$ is convex, and by the linearity the constraint set $\mathcal{P}_{[0,n]}(\kappa)$ expressed in $\mathbf{P}^{C1}(da^n|b^{n-1})$, is convex. Hence, if there exists a maximizing distribution and the so-called Slater condition holds (i.e., a sufficient condition is the existence of an interior point to the constraint set), then the constraint and unconstraint problems are equivalent. For finite alphabet spaces all conditions are easily checked.

Next, the variational equality of Theorem II.1 is applied, together with stochastic optimal control techniques, to provide an

alternative derivation of Theorem III.1, through upper bounds which are achievable, over the smaller set of channel input conditional distributions $\overline{\mathcal{P}}_{[0,n]}^A$. The next theorem is simply introduced to illustrate the importance of the variational equality of directed information in extremum problems of feedback capacity, and to illustrate its importance, when considering channels with limited memory on past channel output symbols.

Theorem III.2. (*Derivation of Theorem III.1 via variational equality*)

Consider the extremum problem of Class A channels defined by (III.67) (which is investigated in Theorem III.1, Part A).

Let $\{V_i(\cdot|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i = 0, \dots, n\}$ be a sequence of conditions distributions on $\{\mathbb{B}_i : i = 0, \dots, n\}$, not necessarily the one generated by the channel and channel input conditional distributions.

Then

$$C_{A^n \rightarrow B^n}^{FB,A} = \sup_{\{\pi_i(da_i|b^{i-1}) : i=0, \dots, n\}} \inf_{\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i=0, \dots, n\}} \sum_{i=0}^n \mathbf{E}^\pi \left\{ \log \left(\frac{dQ_i(\cdot|B^{i-1}, A_i)}{dV_i(\cdot|B^{i-1})}(B_i) \right) \right\} \quad (\text{III.99})$$

$$= \sup_{\{\pi_i(da_i|b^{i-1}) \in \mathcal{M}(\mathbb{A}_i) : i=0, \dots, n\}} \sum_{i=0}^n \mathbf{E}^\pi \left\{ \log \left(\frac{dQ_i(\cdot|B^{i-1}, A_i)}{dV_i^*(\cdot|B^{i-1})}(B_i) \right) \right\}, i = 0, \dots, n \quad (\text{III.100})$$

where $\{V_i^*(\cdot|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i = 0, \dots, n\}$ is given by

$$V_i^*(db_i|b^{i-1}) \triangleq \Pi_i^\pi(db_i|b^{i-1}) = \int_{\mathbb{A}_i} Q_i(db_i|b^{i-1}, a_i) \otimes \pi_i(da_i|b^{i-1}), \quad i = 0, 1, \dots, n. \quad (\text{III.101})$$

Proof: By the variational equality of Theorem II.1, for any arbitrary conditional distribution, $V_i(\cdot|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i), i = 0, 1, \dots, n$, it can be shown that the following identity holds.

$$\sup_{\{P_i(da_i|a^{i-1}, b^{i-1}) : i=0, \dots, n\} \in \mathcal{P}_{[0,n]}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b^{i-1}, a_i)}{d\Pi_i^P(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^P(db_i, db^{i-1}, da_i) \quad (\text{III.102})$$

$$= \sup_{\{P_i(da_i|a^{i-1}, b^{i-1}) : i=0, \dots, n\} \in \mathcal{P}_{[0,n]}} \inf_{\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i=0, \dots, n\}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b^{i-1}, a_i)}{dV_i(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^P(b_i, b^{i-1}, a_i) \quad (\text{III.103})$$

$$\leq \sup_{\{P_i(da_i|a^{i-1}, b^{i-1}) : i=0, \dots, n\} \in \mathcal{P}_{[0,n]}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b^{i-1}, a_i)}{dV_i(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^P(b_i, b^{i-1}, a_i), \quad \forall V_i(db_i|b^{i-1}), i = 0, \dots, n \quad (\text{III.104})$$

where the last inequality holds for any arbitrary distribution $V_i(\cdot|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i), i = 0, 1, \dots, n$, not necessarily the one generated by $\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, 1, \dots, n\} \in \mathcal{P}_{[0,n]}$ and the channel distribution.

Next, define the pay-off function

$$\ell(a_i, b^{i-1}) \triangleq \int_{\mathbb{B}_i} \log \left(\frac{dQ_i(\cdot|b_{i-1}, a_i)}{dV_i(\cdot|b^{i-1})}(b_i) \right) Q_i(db_i|b^{i-1}, a_i) \equiv \ell(a_i, s_i), \quad s_i = b^{i-1}, \quad i = 0, \dots, n \quad (\text{III.105})$$

Since $\mathbf{P}^P(db_i, db^{i-1}, da_i) = Q_i(db_i|db^{i-1}, a_i) \otimes \mathbf{P}(da_i|b^{i-1}) \otimes \mathbf{P}^P(db^{i-1})$, for any arbitrary $V_i(\cdot|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i), i = 0, 1, \dots, n$, by maximizing the right hand side of (III.104) the following upper bound is obtained.

$$(III.102) \leq \sup_{\{P_i(da_i|a^{i-1}, b^{i-1}) : i=0, \dots, n\} \in \mathcal{P}_{[0,n]}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-1}, a_i)}{dV_i(\cdot|b^{i-1})}(b_i) \right) Q_i(db_i|db^{i-1}, a_i) \otimes \mathbf{P}(da_i|b^{i-1}) \otimes \mathbf{P}^P(db^{i-1}) \quad (\text{III.106})$$

$$= \sup_{\{P_i(da_i|a^{i-1}, b^{i-1}) : i=0, \dots, n\} \in \mathcal{P}_{[0,n]}} \sum_{i=0}^n \mathbf{E}^P \left\{ \ell(A_i, B^{i-1}) \right\} \quad (\text{III.107})$$

$$\stackrel{(\alpha)}{=} \sup_{\{\pi_i(da_i|b^{i-1}) \in \mathcal{M}(\mathbb{A}_i) : i=0, \dots, n\}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-1}, a_i)}{dV_i(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^\pi(db_i, db^{i-1}, da_i), \quad \forall V_i(db_i|b^{i-1}), i = 0, \dots, n \quad (\text{III.108})$$

where $\mathbf{P}^\pi(db_i, db^{i-1}, da_i) = Q_i(db_i|db^{i-1}, a_i) \otimes \pi_i(da_i|b^{i-1}) \otimes \mathbf{P}^\pi(db^{i-1}), i = 0, \dots, n$, and the equality in (α) is obtained as follows. Since for each i , the pay-off over which the expectation is taken in (III.107) is $\ell(a_i, b^{i-1}) \equiv \ell(a_i, s_i)$ and $\{S_i : i = 0, \dots, n\}$ is

Markov, as shown in the proof of Theorem III.1, then the maximization occurs in the subset satisfying conditional independence $P_i(da_i|a^{i-1}, s_i) = \mathbf{P}(da_i|s_i) \equiv \pi_i(da_i|s_i) - a.a.(a^{i-1}, s_i), i = 0, \dots, n$, hence $\mathbf{P}^P(db_i, db^{i-1}, da_i) = \mathbf{P}^\pi(db_i, db^{i-1}, da_i), i = 0, \dots, n$, and (III.108) is obtained.

Since the distribution $\{V_i(db_i|b^{i-1}) : i = 0, \dots, n\}$ is arbitrary, by letting this to be the one defined by the channel and $\{\pi_i(da_i|b^{i-1}) \in \mathcal{M}(\mathbb{A}_i) : i = 0, \dots, n\}$, given by (III.101), i.e., $V_i(db_i|b^{i-1}) \triangleq \Pi_i^\pi(db_i|b^{i-1}), i = 0, \dots, n$, then the following upper bound holds.

$$(III.102) \leq \sup_{\{\pi_i(da_i|b^{i-1}) \in \mathcal{M}(\mathbb{A}_i) : i=0,1,\dots,n\}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-1}, a_i)}{d\Pi_i^\pi(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^\pi(db_i, db^{i-1}, da_i) \quad (III.109)$$

Next, it is shown, that the reverse inequality in (III.109) holds, thus establishing the claim. Recall definition (III.65) of $\overline{\mathcal{P}}_{[0,n]}^A$. Since $\overline{\mathcal{P}}_{[0,n]}^A \subset \mathcal{P}_{[0,n]}$, it can be shown that the following inequality holds.

$$\sup_{\{\pi_i(da_i|b^{i-1}) \in \mathcal{M}(\mathbb{A}_i) : i=0,\dots,n\}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b^{i-1}, a_i)}{d\Pi_i^{\pi_i}(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^\pi(db_i, db^{i-1}, da_i) \quad (III.110)$$

$$\leq \sup_{\{P_i(da_i|a^{i-1}, b^{i-1}) : i=0,\dots,n\} \in \mathcal{P}_{[0,n]}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b^{i-1}, a_i)}{d\Pi_i^P(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^P(db_i, db^{i-1}, da_i) = (III.103) \quad (III.111)$$

where $\{\Pi_i^P(db_i|b^{i-1}), \Pi_i^{\pi_i}(db_i|b^{i-1}) : i = 0, \dots, n\}$ are defined by (III.69), (III.70), and $\{\mathbf{P}^P(db_i, db^{i-1}, da_i) : i = 0, \dots, n\}$, $\{\mathbf{P}^\pi(db_i, db^{i-1}, da_i) : i = 0, \dots, n\}$ are induced by the channel and $\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, \dots, n\}$, $\{\pi_i(db_i|b_{i-1}) : i = 0, \dots, n\}$, respectively.

Combining inequalities (III.111) and (III.109) establishes the equality in (III.100), under (III.101). \blacksquare

Note that Theorem III.2 can be used to derive Theorem III.1, Part B, by repeating the above derivation, with the supremum over the set $\mathcal{P}_{[0,n]}$ replaced by the set $\mathcal{P}_{[0,n]}(\kappa)$ in all equations.

B. Channels Class B and Transmission Costs Class A or B

In this section, the information structure of channel input distributions, which maximize $I(A^n \rightarrow B^n)$ is derived for channel distributions of Class B and transmission cost functions Class A or B. The derivation is based on applying the results of Section III-A, and the variational equality of directed information, to show the supremum over all channel input conditional distributions occurs in a smaller set $\overset{\circ}{\mathcal{P}}_{[0,n]} \subseteq \overline{\mathcal{P}}_{[0,n]}^A \subset \mathcal{P}_{[0,n]}$ (for Class B transmission costs the subset is strictly smaller, i.e., $\overset{\circ}{\mathcal{P}}_{[0,n]} \subset \overline{\mathcal{P}}_{[0,n]}^A$).

The derivation is first presented for any channel distribution of Class B and transmission cost of Class B, with $M = 2, K = 1$, to illustrate the procedure, as the derivation of the general cases are similar.

1) Channel Class B and Transmission Cost Class B, $M = 2, K = 1$: First, consider any channel distribution of Class B, with $M = 2$, i.e., $Q_i(db_i|b_{i-1}, b_{i-2}, a_i), i = 0, 1, \dots, n$, without transmission cost.

Then the FTFI capacity is defined by

$$C_{A^n \rightarrow B^n}^{FB,B,2} \triangleq \sup_{\{P_i(da_i|a^{i-1}, b^{i-1}) : i=0,\dots,n\} \in \mathcal{P}_{[0,n]}} \sum_{i=0}^n \mathbf{E}^P \left\{ \log \left(\frac{dQ_i(\cdot|B_{i-1}, B_{i-2}, A_i)}{d\Pi_i^P(\cdot|B^{i-1})}(B_i) \right) \right\} \quad (III.112)$$

$$\stackrel{(\alpha)}{=} \sup_{\{\pi_i(da_i|b^{i-1}) \in \mathcal{M}(\mathbb{A}_i) : i=0,\dots,n\}} \sum_{i=0}^n \mathbf{E}^\pi \left\{ \log \left(\frac{dQ_i(\cdot|B_{i-1}, B_{i-2}, A_i)}{d\Pi_i^{\pi_i}(\cdot|B^{i-1})}(B_i) \right) \right\} \quad (III.113)$$

$$\Pi_i^{\pi_i}(db_i|b^{i-1}) = \int_{\mathbb{A}_i} Q_i(db_i|b_{i-1}, b_{i-2}, a_i) \otimes \pi_i(da_i|b^{i-1}), \quad i = 0, 1, \dots, n \quad (III.114)$$

where (α) is due to Theorem III.1, because the set of channel distributions of Class B is a subset of the set of channel distributions of Class A, and the joint distribution over which $\mathbf{E}^\pi\{\cdot\}$ is taken is $\mathbf{P}^\pi(da_i, db^i) \equiv \mathbf{P}^{\pi_0, \pi_1, \dots, \pi_i}(da_i, db^i), i = 0, 1, \dots, n$.

The main challenge is to show the optimal channel input distribution induces the following conditional independence on the transition probability of the channel output process: $\mathbf{P}(db_i|b^{i-1}) = \mathbf{P}(db_i|b_{i-1}, b_{i-2}) - a.a. b^{i-1}, i = 0, \dots, n$.

This is shown by invoking, *Step 2*, of the two-step procedure (i.e., the variational equality of directed information), to deduce that the maximization in (III.113) occurs in $\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2} \triangleq \{\pi_i^2(da_i|b_{i-1}, b_{i-2}) : i = 0, \dots, n\} \subset \overline{\mathcal{P}}_{[0,n]}^A \triangleq \{\pi_i(da_i|b^{i-1}) : i = 0, \dots, n\}$, and that $\Pi_i^{\pi_i}(db_i|b^{i-1}) = v_i^{\pi_i^2}(db_i|b_{i-1}, b_{i-2}) - a.a. b^{i-1}, i = 0, \dots, n$, that is, the optimal channel input distribution satisfies conditional independence property, $\pi_i(da_i|b^{i-1}) = \pi_i^2(da_i|b_{i-1}, b_{i-2}) - a.a. b^{i-1}, i = 0, \dots, n$.

Lemma III.1. (*Characterization of FTFI capacity for channels of class B and transmission costs of class B, $M = 2, K = 1$*)

Suppose the channel distribution is of Class B with $M = 2$.

Define the restricted class of policies $\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2} \subset \overline{\mathcal{P}}_{[0,n]}^A$ by

$$\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2} \triangleq \left\{ \left\{ \pi_i(da_i|b^{i-1}) : i = 0, 1, \dots, n \right\} \in \overline{\mathcal{P}}_{[0,n]}^A : \pi_i(da_i|b^{i-1}) = \pi_i^2(da_i|b_{i-1}, b_{i-2}) - a.a. b^{i-1}, i = 0, 1, \dots, n \right\}.$$

Then the following hold.

Part A. The maximization in (III.112) over $\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, \dots, n\} \in \overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2}$ occurs in the smaller class $\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2}$, that is, it satisfies the following conditional independence.

$$P_i(da_i|a^{i-1}, b^{i-1}) = \pi_i^2(da_i|b_{i-1}, b_{i-2}) - a.a. (a^{i-1}, b^{i-1}), \quad i = 0, 1, \dots, n. \quad (\text{III.115})$$

Moreover, any distribution from the class $\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2}$ induces a channel output process $\{B_i : i = 0, 1, \dots, n\}$, with conditional probabilities which are second-order Markov, that is,

$$\Pi_i^P(db_i|b^{i-1}) = v_i^{\pi_i^2}(db_i|b_{i-1}, b_{i-2}) - a.a. b^{i-1}, \quad i = 0, 1, \dots, n, \quad (\text{III.116})$$

and the characterization of FTFI capacity is given by the following expression.

$$C_{A^n \rightarrow B^n}^{FB,B,2} = \sup_{\{P_i^M(da_i|b_{i-1}, b_{i-2}) \in \mathcal{M}(\mathbb{A}_i), i=0, \dots, n\}} \sum_{i=0}^n \mathbf{E}^{\pi^2} \left\{ \log \left(\frac{dQ_i(\cdot|B_{i-1}, B_{i-2}, A_i)}{dv_i^{\pi_i^2}(\cdot|B_{i-1}, B_{i-2})} (B_i) \right) \right\} \quad (\text{III.117})$$

$$\equiv \sup_{\{\pi_i^2(da_i|b_{i-1}, b_{i-2}) \in \mathcal{M}(\mathbb{A}_i), i=0, \dots, n\}} \sum_{i=0}^n I(A_i; B_i | B_{i-1}, B_{i-2}) \quad (\text{III.118})$$

where

$$v_i^{\pi_i^2}(db_i|b_{i-1}, b_{i-2}) = \int_{\mathbb{A}_i} Q_i(db_i|b_{i-1}, b_{i-2}, a_i) \otimes \pi_i^2(da_i|b_{i-1}, b_{i-2}), \quad i = 0, 1, \dots, n. \quad (\text{III.119})$$

Part B. Suppose the following two conditions hold.

$$(a) \quad \gamma_i(T^i a^n, T^i b^{n-1}) = \gamma_i^{B,1}(a_i, b_{i-1}), \quad i = 0, \dots, n, \quad (\text{III.120})$$

$$(b) \quad C_{A^n \rightarrow B^n}^{FB,B,1}(\kappa) \triangleq \sup_{\{P_i(da_i|a^{i-1}, b^{i-1}) : i=0, \dots, n\} \in \overset{\circ}{\mathcal{P}}_{[0,n]}(\kappa)} \sum_{i=0}^n \mathbf{E}^P \left\{ \log \left(\frac{dQ_i(\cdot|B_{i-1}, B_{i-2}, A_i)}{d\Pi_i^P(\cdot|B^{i-1})} (B_i) \right) \right\} \quad (\text{III.121})$$

$$= \inf_{s \geq 0} \sup_{\{P_j(\cdot|a^{j-1}, b^{j-1}) : j=0, \dots, n\} \in \overset{\circ}{\mathcal{P}}_{[0,n]}(\kappa)} \left\{ I(A^n \rightarrow B^n) - s \left\{ \mathbf{E}^P \left(c_{0,n}(A^n, B^{n-1}) \right) - \kappa(n+1) \right\} \right\}. \quad (\text{III.122})$$

The characterization of FTFI capacity is given by the following expression.

$$C_{A^n \rightarrow B^n}^{FB,B,2}(\kappa) = \sup_{\{\pi_i^2(da_i|b_{i-1}, b_{i-2}) \in \mathcal{M}(\mathbb{A}_i), i=0, \dots, n; \frac{1}{n+1} \mathbf{E}^{\pi^2} \left\{ \sum_{i=0}^n \gamma_i^{B,1}(A_i, B_{i-1}) \right\} \leq \kappa\}} \sum_{i=0}^n I(A_i; B_i | B_{i-1}, B_{i-2}) \quad (\text{III.123})$$

That is, in Part A, B the conditional distribution of the joint process $\{(A_i, B_i) : i = 0, 1, \dots, n\}$ satisfies (I.28) and the channel output process $\{B_i : i = 0, 1, \dots, n\}$ is a second-order Markov process, i.e., its conditional distribution satisfies (I.29).

Proof: Part A. Since the channel distribution of Class B, with $M = 2$, is a special case the channel distribution (I.15), the

statements of Theorem III.1 hold, hence (III.112)-(III.114) hold, and the maximization over $\mathcal{P}_{[0,n]}$ occurs in the preliminary set $\overline{\mathcal{P}}_{[0,n]}^A$ defined by (III.65). By (III.113), and since $\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2} \subset \overline{\mathcal{P}}_{[0,n]}^A$, then

$$C_{A^n \rightarrow B^n}^{FB,B,2} = \sup_{\{\pi_i(da_i|b^{i-1}): i=0,\dots,n\}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-1}, b_{i-2}, a_i)}{d\Pi_i^\pi(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^\pi(db_i, db^{i-1}, da_i) \quad (\text{III.124})$$

$$\geq \sup_{\{\pi_i(da_i|b^{i-1}): i=0,\dots,n\} \in \overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-1}, b_{i-2}, a_i)}{d\Pi_i^\pi(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^\pi(db_i, db^{i-1}, da_i) \quad (\text{III.125})$$

$$\geq \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-1}, b_{i-2}, a_i)}{d\nu_i^{\pi^2}(\cdot|b_{i-1}, b_{i-2})}(b_i) \right) \mathbf{P}^{\pi^2}(db_i, db_{i-1}, b_{i-2}, da_i), \quad \forall \pi_i^2(da_i|b_{i-1}, b_{i-2}) \in \mathcal{M}(\mathbb{A}_i), i=0, \dots, n \quad (\text{III.126})$$

where $\{\Pi_i^\pi(db_i|b^{i-1}), \nu_i^{\pi^2}(db_i|b_{i-1}, b_{i-2}) : i=0, \dots, n\}$ are given by (III.114), (III.119), and $\{\mathbf{P}^\pi(db_i, db^{i-1}, da_i), \mathbf{P}^{\pi^2}(db_i, db_{i-1}, b_{i-2}, da_i) : i=0, \dots, n\}$ are induced by the channel and $\{\pi_i(da_i|b^{i-1}), \pi_i^2(da_i|b_{i-1}, b_{i-2}) : i=0, \dots, n\}$, respectively. Taking the supremum over $\{\pi_i^2(da_i|b_{i-1}, b_{i-2}) : i=0, \dots, n\}$, inequality (III.126) is retained and hence, the following lower bound is obtained.

$$C_{A^n \rightarrow B^n}^{FB,B,2} = (\text{III.124}) \geq \sup_{\{\pi_i^2(da_i|b_{i-1}, b_{i-2}): i=0,\dots,n\}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-1}, b_{i-2}, a_i)}{d\nu_i^{\pi^2}(\cdot|b_{i-1}, b_{i-2})}(b_i) \right) \mathbf{P}^{\pi^2}(db_i, db_{i-1}, b_{i-2}, da_i) \quad (\text{III.127})$$

$$\equiv \sup_{\{\pi_i^2(da_i|b_{i-1}, b_{i-2}): i=0,\dots,n\}} \sum_{i=0}^n I(A_i; B_i | B_{i-1}, B_{i-2}). \quad (\text{III.128})$$

Next, the variational equality of Theorem II.1 is applied to show the reverse inequality in (III.127) holds. Given a policy from the set $\overline{\mathcal{P}}_{[0,n]}^A$, and any arbitrary distribution $\{V_i(db_i|b^{i-1}) : i=0, \dots, n\} \in \mathcal{M}(\mathbb{B}_i) : i=0, \dots, n\}$, then

$$C_{A^n \rightarrow B^n}^{FB,B,2} = \sup_{\{\pi_i(da_i|b^{i-1}): i=0,\dots,n\}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-1}, b_{i-2}, a_i)}{d\Pi_i^\pi(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^\pi(db_i, db^{i-1}, da_i) \quad (\text{III.129})$$

$$= \sup_{\{\pi_i(da_i|b^{i-1}): i=0,\dots,n\}} \inf_{\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i=0,\dots,n\}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-1}, b_{i-2}, a_i)}{dV_i(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^\pi(db_i, db^{i-1}, da_i) \quad (\text{III.130})$$

where $\{\mathbf{P}^\pi(b_i, b^{i-1}, da_i) : i=0, \dots, n\}$ is defined by the channel distribution and $\{\pi_i(da_i|b^{i-1}) : i=0, 1, \dots, n\} \in \overline{\mathcal{P}}_{[0,n]}^A$. Since $\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i=0, \dots, n\}$ is arbitrary, then an upper bound for (III.130) is obtained as follows. Assume the arbitrary channel output conditional probability is the one satisfying the conditional independence

$$V_i(db_i|b^{i-1}) = \bar{V}_i(db_i|b_{i-1}, b_{i-2}) - a.a. b^{i-1}, \quad i=0, 1, \dots, n. \quad (\text{III.131})$$

Define the pay-off

$$\ell_i(a_i, s_i) \triangleq \int_{\mathbb{B}_i} \log \left(\frac{dQ_i(\cdot|s_i, a_i)}{d\bar{V}_i(\cdot|s_i)}(b_i) \right) Q_i(db_i|s_i, a_i), \quad s_i \triangleq (b_{i-1}, b_{i-2}), \quad i=0, \dots, n. \quad (\text{III.132})$$

Then, by removing the infimum in (III.130) over $\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i=0, \dots, n\}$, and substituting (III.131), the following

upper bound is obtained.

$$C_{A^n \rightarrow B^n}^{FB,B,2} \leq \sup_{\{\pi_i(da_i|b^{i-1}): i=0, \dots, n\}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-1}, b_{i-2}, a_i)}{d\bar{V}_i(\cdot|b_{i-1}, b_{i-2})}(b_i) \right) \mathbf{P}^\pi(db_i, db_{i-1}, db_{i-2}, da_i), \quad \forall \bar{V}_i(db_i|b_{i-1}, b_{i-2}), i=0, \dots, n \quad (\text{III.133})$$

$$\stackrel{(\alpha)}{=} \sup_{\{\pi_i(da_i|b^{i-1}): i=0, \dots, n\}} \sum_{i=0}^n \mathbf{E}^\pi \left\{ \int_{\mathbb{B}_i} \log \left(\frac{dQ_i(\cdot|b_{i-1}, b_{i-2}, a_i)}{d\bar{V}_i(\cdot|b_{i-1}, b_{i-2})}(b_i) \right) Q_i(db_i|b_{i-1}, b_{i-2}, a_i) \right\}, \quad \forall \bar{V}_i(db_i|b_{i-1}, b_{i-2}), i=0, \dots, n \quad (\text{III.134})$$

$$\equiv \sup_{\{\pi_i(da_i|b^{i-1}): i=0, \dots, n\}} \sum_{i=0}^n \mathbf{E}^\pi \left\{ \ell_i(A_i, S_i) \right\}, \quad \forall \bar{V}_i(db_i|s_i), \quad i=0, \dots, n \quad (\text{III.135})$$

$$\stackrel{(\beta)}{=} \sup_{\{\pi_i^2(da_i|b_{i-1}, b_{i-2}): i=0, \dots, n\}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-1}, b_{i-2}, a_i)}{d\bar{V}_i(\cdot|b_{i-1}, b_{i-2})}(b_i) \right) \mathbf{P}^{\pi^2}(db_i, db_{i-1}, db_{i-2}, da_i), \quad \forall \bar{V}_i(db_i|b_{i-1}, b_{i-2}), i=0, \dots, n \quad (\text{III.136})$$

where (α) is by definition, and (β) is obtained as follows. Since the pay-off in (III.135), i.e., $\ell_i(\cdot, \cdot)$ is a function of (a_i, s_i) , for $i=0, \dots, n$, then $\{S_i : i=0, \dots, n\}$ is the state process controlled by $\{A_i : i=0, \dots, n\}$. Moreover, by virtue of Bayes' theorem, and the channel definition, the following identity holds.

$$\mathbf{P}(ds_{i+1}|s^i, a^i) = \mathbf{P}(ds_{i+1}|s_i, a_i), \quad i=0, \dots, n-1. \quad (\text{III.137})$$

In view of the Markov structure of the controlled process $\{S_i : i=0, \dots, n\}$, i.e., (III.137), then the expectation of the pay-off in (III.135) is given by

$$\sum_{i=0}^n \mathbf{E}^\pi \left\{ \ell(A_i, S_i) \right\} = \int_{\mathbb{B}_{i-1} \times \mathbb{B}_{i-2} \times \mathbb{A}_i} \ell(a_i, s_i) \mathbf{P}(da_i|s_i) \mathbf{P}^\pi(ds_i), \quad \forall \bar{V}_i(db_i|s_i), \quad i=0, \dots, n. \quad (\text{III.138})$$

Thus, $\{\mathbf{P}(ds_{i+1}|s_i, a_i), \quad i=0, \dots, n-1\}$ is the controlled object and by the discussion on classical stochastic optimal control, i.e., Feature 1, the supremum over $\{\pi_i(da_i|b^{i-1}) : i=0, \dots, n\}$ in (III.133), satisfies $\pi_i(da_i|b^{i-1}) = \pi_i^2(da_i|b_{i-1}, b_{i-2}) - a.a.b^{i-1}, i=0, \dots, n$.

Alternatively, this is shown directly as follows. Note that

$$\mathbf{P}^\pi(ds_i) = \int \mathbf{P}(ds_i|s_{i-1}, a_{i-1}) \mathbf{P}(da_{i-1}|s_{i-1}) \mathbf{P}^\pi(ds_{i-1}) \implies \mathbf{P}^{\pi^2}(ds_i) = \int \mathbf{P}(ds_i|s_{i-1}, a_{i-1}) \mathbf{P}(da_{i-1}|s_{i-1}) \mathbf{P}^{\pi^2}(ds_{i-1}) \quad (\text{III.139})$$

that is, $\mathbf{P}^\pi(ds_i) \equiv \mathbf{P}^{\pi^2}(ds_i)$, depends on $\{\mathbf{P}(da_j|s_j) \equiv \pi_j^2(da_j|s_j) : j=0, \dots, i-1\}$, and hence the right hand side in (III.138) depends on $\{\mathbf{P}(da_j|s_j) \equiv \pi_j^2(da_j|s_j) : j=0, \dots, i\}$. This implies, the supremum over $\{\pi_i(da_i|b^{i-1}) : i=0, \dots, n\}$ in (III.133), satisfies $\pi_i(da_i|b^{i-1}) = \pi_i^2(da_i|b_{i-1}, b_{i-2}) - a.a.b^{i-1}, i=0, \dots, n$, that is, the controlled object is second-order Markov, and consequently, $\mathbf{P}^\pi(db_i, db_{i-1}, db_{i-2}, da_i) = Q_i(db_i|b_{i-1}, b_{i-2}, a_i) \otimes \pi_i^2(da_i|b_{i-1}, b_{i-2}) \otimes \mathbf{P}^{\pi^2}(db_{i-1}, db_{i-2}) \equiv \mathbf{P}^{\pi^2}(db_i, db_{i-1}, db_{i-2}, da_i), i=0, \dots, n$. Hence, (III.136) is obtained.

Since $\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i=0, \dots, n\}$ satisfying (III.131), is an arbitrary distribution, let

$$\begin{aligned} V_i(db_i|b^{i-1}) &\triangleq \int_{\mathbb{A}_i} Q_i(db_i|b_{i-1}, b_{i-2}, a_i) \otimes \mathbf{P}^\pi(da_i|b^{i-1}) \equiv V_i^\pi(db_i|b^{i-1}) \\ &= \bar{V}_i(db_i|b_{i-1}, b_{i-2}) \triangleq \int_{\mathbb{A}_i} Q_i(db_i|b_{i-1}, b_{i-2}, a_i) \otimes \pi_i^2(da_i|b_{i-1}, b_{i-2}) = \bar{V}_i^{\pi^2}(db_i|b_{i-1}, b_{i-2}) \end{aligned} \quad (\text{III.140})$$

$$\equiv v_i^{\pi^2}(db_i|b_{i-1}, b_{i-2}) - a.a.b^{i-1}, \quad i=0, 1, \dots, n. \quad (\text{III.141})$$

Then by substituting (III.141) into (III.136), the following inequality is obtained.

$$C_{A^n \rightarrow B^n}^{FB,B,2} \leq \sup_{\{\pi_i^2(da_i|b_{i-1}, b_{i-2}): i=0, \dots, n\}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-1}, b_{i-2}, a_i)}{dv_i^{\pi^2}(\cdot|b_{i-1}, b_{i-2})}(b_i) \right) \mathbf{P}^{\pi^2}(db_i, db_{i-1}, db_{i-2}, da_i) \quad (\text{III.142})$$

$$\equiv \sup_{\{\pi_i^2(da_i|b_{i-1}, b_{i-2}): i=0, \dots, n\}} \sum_{i=0}^n I(A_i; B_i|B_{i-1}, B_{i-2}) \quad (\text{III.143})$$

Combining (III.128) and (III.143), the supremum over $\{\pi_i(da_i|b^{i-1}) : i = 0, \dots, n\}$ in $C_{A^n \rightarrow B^n}^{FB,B,2}$ occurs in the subset $\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,2}$, and hence (III.115)-(III.119) are a consequence of this fact. Part B. Using the definition of the transmission cost function (III.120) and (III.137), then

$$\sum_{i=0}^n \mathbf{E}^P \left\{ \gamma_i^{B,1}(A_i, B_{i-1}) \right\} = \sum_{i=0}^n \mathbf{E}^\pi \left\{ \gamma_i^{B,1}(A_i, B_{i-1}) \right\} = \sum_{i=0}^n \mathbf{E}^{\pi^2} \left\{ \int_{\mathbb{A}_i} \gamma_i^{B,1}(a_i, B_{i-1}) \pi^2(a_i|B_{i-1}) \right\} \quad (\text{III.144})$$

where the last equality is due to $S_i = (B_{i-1}, B_i - 2), i = 0, \dots, n$ and the sample path pay-off depends only on B_{i-1} . The above expectation is a function of $\{\pi_i^2(da_i|b_{i-1}, b_{i-2}) : i = 0, \dots, n\}$, hence by Part A, the results are obtained. This completes the prove. ■

The following remark clarifies certain aspects of the application of variational equality.

Remark III.3. (On the application of variational equality in Lemma III.1)

(a) The important point to be made regarding Lemma III.1 is that, for any channel of Class B with $M = 2$ and transmission cost of class B with $K = 1$, the information structure of channel input conditional distribution, which maximizes directed information $I(A^n \rightarrow B^n)$ is $\mathcal{J}_i^P = \{b_{i-1}, b_{i-2}\}, i = 0, 1, \dots, n$, and it is determined by $\max\{K, M\}$.

(b) From Lemma III.1, it follows that if the channel is replaced by $\{Q_i(db_i|b_{i-1}, a_i) : i = 0, \dots, n\}$, the information structure of channel input conditional distribution, which maximizes directed information $I(A^n \rightarrow B^n)$ is $\mathcal{J}_i^P = \{b_{i-1}\}, i = 0, 1, \dots, n$, and the corresponding characterization of FTFI capacity is

$$C_{A^n \rightarrow B^n}^{FB,B,1} = \sup_{\{\pi_i^M(da_i|b_{i-1}) \in \mathcal{M}(\mathbb{A}_i), i=0, \dots, n\}} \sum_{i=0}^n I(A_i; B_i|B_{i-1}). \quad (\text{III.145})$$

(c) By Lemma III.1, if the channel is memoryless (i.e., $M = 0$), and the transmission cost constraint is $\frac{1}{n+1} \mathbf{E} \left\{ \sum_{i=0}^n \gamma_i^{B,1}(A_i, B_{i-1}) \right\} \leq \kappa$, then the information structure corresponding to the channel input conditional distribution, which maximizes directed information $I(A^n \rightarrow B^n)$, is $\mathcal{J}_i^P = \{b_{i-1}\}, i = 0, 1, \dots, n$, and the corresponding characterization of FTFI capacity is

$$C_{A^n \rightarrow B^n}^{FB,B,1}(\kappa) = \sup_{\pi_i^1(da_i|b_{i-1}) \in \mathcal{M}(\mathbb{A}_i), i=0, \dots, n; \frac{1}{n+1} \mathbf{E}^{\pi^1} \left\{ \sum_{i=0}^n \gamma_i^{B,1}(A_i, B_{i-1}) \right\} \leq \kappa} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|a_i)}{dv_i^{\pi^1}(\cdot|b_{i-1})}(b_i) \right) \mathbf{P}^{\pi^1}(db_i, db_{i-1}, da_i) \quad (\text{III.146})$$

where

$$v_i^{\pi^1}(db_i|b_{i-1}) = \int_{\mathbb{A}_i} Q_i(db_i|a_i) \otimes \pi_i^1(da_i|b_{i-1}), \quad i = 0, 1, \dots, n. \quad (\text{III.147})$$

(d) Memoryless Channels. If the transmission cost constraint in (c) is replaced by $\frac{1}{n+1} \mathbf{E} \left\{ \sum_{i=0}^n \gamma_i^{B,0}(A_i) \right\} \leq \kappa$, since the channel is memoryless, then the derivation of Lemma III.1 can be repeated with (III.131) replaced by $V_i(db_i|b^{i-1}) = \bar{V}_i(db_i) - a.a.b^{i-1}, i = 0, 1, \dots, n$, to deduce that in all equations in (c), the optimal channel input distribution $\pi_i^1(da_i|b_{i-1})$ is replaced by $\pi_i(da_i), i = 0, \dots, n$, and $C_{A^n \rightarrow B^n}^{FB,B,1}(\kappa) = C_{A^n \rightarrow B^n}^{FB,B,0}(\kappa) = \sup_{\pi(da_i): i=0, \dots, n} \sum_{i=0}^n I(A_i; B_i)$, as expected. That is, it is possible to derive the capacity achieving conditional independence property of memoryless channels with feedback, directly, without first showing via the converse to the coding theorem that feedback does not increase capacity (see [10]).

(e) The derivation of Theorem III.1 is easily extended to any channel of Class B and transmission cost function of Class B (i.e., with M, K arbitrary); this is done next.

2) Channels Class B and Transmission Costs Class A or B: Consider any channel distribution of Class B defined by (I.16), i.e., given by $\{Q_i(db_i|b_{i-M}^{i-1}, a_i) : i = 0, 1, \dots, n\}$.

Since the induced joint distribution is $\mathbf{P}^P(da^i, db^i) = \otimes_{j=0}^i P_j(da_j|a^{j-1}, b^{j-1}) \otimes Q_j(db_j|b_{j-M}^{j-1}, a_j), i = 0, \dots, n$, the FTFI capacity is defined by

$$C_{A^n \rightarrow B^n}^{FB,B,M} \triangleq \sup_{\{P_i(da_i|a^{i-1}, b^{i-1}): i=0, \dots, n\} \in \mathcal{P}_{[0,n]}} \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-M}^{i-1}, a_i)}{d\mathbf{P}_i^P(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^P(db_{i-M}^i, da_i) \quad (\text{III.148})$$

where

$$\Pi_i^P(db_i|b^{i-1}) = \int_{\mathbb{A}_i} Q_i(db_i|b_{i-M}^{i-1}, a_i) \otimes P_i(da_i|a^{i-1}, b^{i-1}) \otimes \mathbf{P}^P(da^{i-1}|b^{i-1}), \quad i = 0, \dots, n. \quad (\text{III.149})$$

The next theorems presents various generalizations of Theorem III.1.

Theorem III.3. (Characterization of FTFI capacity of channel class B and transmission costs of class A or B)

Part A. Suppose the channel distribution is of Class B, that is, $\mathbf{P}_{B_i|B^{i-1}, A^i}(db_i|b^{i-1}, a^i) = Q_i(db_i|\mathcal{J}_i^Q)$, where \mathcal{J}_i^Q is given by

$$\mathcal{J}_i^Q = \{b_{i-M}^{i-1}, a_i\}, \quad i = 0, \dots, n \quad (\text{III.150})$$

Then the maximization in (III.148) over $\mathcal{P}_{[0,n]}$ occurs in the subset

$$\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,M} \triangleq \{P_i(da_i|a^{i-1}, b^{i-1}) = \pi_i^M(da_i|b_{i-M}^{i-1}) - a.a.(a^{i-1}, b^{i-1}) : i = 0, 1, \dots, n\} \subset \mathcal{P}_{[0,n]}. \quad (\text{III.151})$$

and the characterization of the FTFI feedback capacity is given by the following expression.

$$C_{A^n \rightarrow B^n}^{FB,B,M} = \sup_{\{\pi_i^M(da_i|b_{i-M}^{i-1}) \in \mathcal{M}(\mathbb{A}_i) : i=0, \dots, n\}} \sum_{i=0}^n \mathbf{E}^{\pi^M} \left\{ \log \left(\frac{dQ_i(\cdot|B_{i-M}^{i-1}, A_i)}{v_i^{\pi^M}(\cdot|B_{i-M}^{i-1})} (B_i) \right) \right\} \quad (\text{III.152})$$

$$\equiv \sup_{\{\pi_i^M(da_i|b_{i-M}^{i-1}) \in \mathcal{M}(\mathbb{A}_i) : i=0, \dots, n\}} \sum_{i=0}^n I(A_i; B_i|B_{i-M}^{i-1}) \quad (\text{III.153})$$

where

$$v_i^{\pi^M}(db_i|b_{i-M}^{i-1}) = \int_{\mathbb{A}_i} Q_i(db_i|b_{i-M}^{i-1}, a_i) \otimes \pi_i^M(da_i|b_{i-M}^{i-1}), \quad i = 0, \dots, n, \quad (\text{III.154})$$

$$\mathbf{P}^{\pi^M}(da_i, db_{i-M}^i) = Q_i(db_i|b_{i-M}^{i-1}, a_i) \otimes \pi_i^M(da_i|b_{i-M}^{i-1}) \otimes \mathbf{P}^{\pi^M}(b_{i-M}^{i-1}), \quad i = 0, \dots, n. \quad (\text{III.155})$$

Part B. Suppose the channel distribution is of Class B as in Part A, and the maximization in (III.148) is over $\mathcal{P}_{0,n}(\kappa)$, defined with respect to transmission cost $\gamma_i(\cdot, \cdot)$, which is measurable with respect to \mathcal{J}_i^γ given by

$$\mathcal{J}_i^\gamma = \{a_i, b_{i-K}^{i-1}\}, \quad i = 0, \dots, n \quad (\text{III.156})$$

and the analogue of Lemma III.1, Part B, (b) holds.

The maximization in (III.148) over $\{P_i(da_i|a^{i-1}, b^{i-1}), i = 0, \dots, n\} \in \mathcal{P}_{0,n}(\kappa)$ occurs in the subset

$$\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,J}(\kappa) \triangleq \left\{ P_i(da_i|a^{i-1}, b^{i-1}) = \pi_i^J(da_i|b_{i-J}^{i-1}) - a.a.(a^{i-1}, b^{i-1}), i = 0, 1, \dots, n : \right. \\ \left. \frac{1}{n+1} \mathbf{E}^{\pi^J} \left(c_{0,n}(A^n, B^{n-1}) \right) \leq \kappa \right\} \subset \mathcal{P}_{[0,n]}(\kappa), \quad J \triangleq \max\{M, K\} \quad (\text{III.157})$$

and the characterization of FTFI capacity is given by the following expression.

$$C_{A^n \rightarrow B^n}^{FB,B,J}(\kappa) = \sup_{\{\pi_i^J(da_i|b_{i-J}^{i-1}) \in \mathcal{M}(\mathbb{A}_i) : i=0, \dots, n\} \in \overset{\circ}{\mathcal{P}}_{[0,n]}^{B,J}(\kappa)} \sum_{i=0}^n \mathbf{E}^{\pi^J} \left\{ \log \left(\frac{dQ_i(\cdot|B_{i-M}^{i-1}, A_i)}{dv_i^{\pi^J}(\cdot|B_{i-J}^{i-1})} (B_i) \right) \right\} \quad (\text{III.158})$$

where

$$\mathbf{P}^{\pi^J}(db_{i-J}^i, da_i) = Q_i(db_i|b_{i-M}^{i-1}, a_i) \otimes \pi_i^J(da_i|b_{i-J}^{i-1}) \otimes \mathbf{P}^{\pi^J}(db_{i-J}^{i-1}), \quad i = 0, 1, \dots, n, \quad (\text{III.159})$$

$$v_i^{\pi^J}(db_i|b_{i-J}^{i-1}) = \int_{\mathbb{A}_i} Q_i(db_i|b_{i-M}^{i-1}, a_i) \otimes \pi_i^J(da_i|b_{i-J}^{i-1}), \quad i = 0, 1, \dots, n. \quad (\text{III.160})$$

Part C. Suppose the channel distribution is of Class B as in Part A, and the maximization in (III.148) is over $\mathcal{P}_{0,n}(\kappa)$, defined with respect to a transmission cost of Class A, $\{\gamma_i^A(a_i, b^{i-1}) : i = 0, \dots, n\}$, and the analogue of Lemma III.1, Part B, (b) holds. The maximization in (III.148) over $\{P_i(da_i|a^{i-1}, b^{i-1}), i = 0, \dots, n\} \in \mathcal{P}_{[0,n]}(\kappa)$ occurs in $\overline{\mathcal{P}}_{[0,n]}^A \cap \mathcal{P}_{[0,n]}$.

Proof: Part A. The derivation is based on the results obtained thus far, using Step 1 and Step 2 of the Two-Step procedure. By Step 2 of the Two-Step Procedure, repeating the derivation of Lemma III.1, if necessary, it can be shown that the optimal channel input distribution occurs in $\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,M}$.

Part B. The case with transmission cost is shown by applying Lagrange duality to define the unconstrained problem, and then noticing that the upper bound resulting from the variational equality of directed information is achievable, provided the arbitrary distribution (analogue of (III.141)) is chosen so that $V_i(db_i|b^{i-1}) = v^{\pi^J}(db_i|b_{i-J}^{i-1}) - a.a.b^{i-1}, i = 0, 1, \dots, n, J \triangleq \max\{M, K\}$, establishing (III.157).

Part C. Since the transmission cost is of Class A, $\{\gamma_i^A(a_i, b^{i-1}) : i = 0, \dots, n\}$, and the Channel distribution is of Class B, the statement of Theorem III.1, Part C holds, hence the set of all channel input distributions, which maximize directed information $I(A^n \rightarrow B^n) = \sum_{i=0}^n \int \log \left(\frac{dQ_i(\cdot|b_{i-M}^{i-1}, a_i)}{d\pi_i^P(\cdot|b^{i-1})}(b_i) \right) \mathbf{P}^P(db_{i-M}, da_i)$, occur in the set $\overset{\circ}{\mathcal{P}}_{[0,n]}^A$. Consequently, the channel output conditional probabilities are given by

$$\Pi_i^P(db_i|b^{i-1}) = \int_{\mathbb{A}^i} Q_i(db_i|b_{i-M}^{i-1}, a_i) \otimes \pi(da_i|b^{i-1}) \equiv \Pi_i^\pi(db_i|b^{i-1}), \quad i = 0, \dots, n \quad (\text{III.161})$$

However, any attempt to apply the variational equality of directed information, as done in Lemma III.1, to derive upper bounds on the corresponding directed information, which are achievable over arbitrary distributions, $\{V_i(db_i|b^{i-1}) \in \mathcal{M}(\mathbb{B}_i) : i = 0, \dots, n\}$, which satisfy conditional independence condition

$$V_i(db_i|b^{i-1}) = \bar{V}_i(db_i|b_{i-L}^{i-1}) - a.a.b^{i-1}, \quad \text{for any finite nonnegative } L, \quad i = 0, 1, \dots, n \quad (\text{III.162})$$

will fail. This is because the transmission cost of Class A, depends, for each i , on the entire past output symbols $\{b^{i-1}\}$, and hence the maximization step, using stochastic optimal control, over channel input distributions from the set $\overset{\circ}{\mathcal{P}}_{[0,n]}^A$, satisfying the average transmission cost constraint cannot occur is a smaller subset, i.e., recall Feature 1 of the discussion on classical stochastic optimal control. This completes the prove. \blacksquare

C. Implications on Dynamic Programming Recursion

In this section, the implications of the information structures of the optimal channel input distributions, are discussed in the context of dynamic programming.

Channels Class B and Transmission Costs Class B. Consider a channel distribution and transmission cost function, both of Class B, given by $\mathbf{P}_{B_i|B^{i-1}, A^i}(db_i|b^{i-1}, a^i) = Q_i(db_i|b_{i-M}^{i-1}, a_i) - a.a.(b^{i-1}, a^i), \gamma_i^{B,K}(a_i, b_{i-K}^{i-1}), i = 0, \dots, n$.

Since the output process $\{B_i : i = 0, \dots, n\}$ is $J = \max\{M, K\}$ -order Markov, i.e., (III.160), holds, and the characterization of FTFI capacity is given by (III.158), the optimization over $\overset{\circ}{\mathcal{P}}_{[0,n]}^{B,J}(\kappa)$ can be solved via dynamic programming, as follows.

Let $C_t^{B,J} : \mathbb{B}_{t-J}^{t-1} \mapsto \mathbb{R}$ denote the cost-to-go corresponding to (III.158) from time “ t ” to the terminal time “ n ” given the values of the output and input $B_{t-J}^{t-1} = b_{t-J}^{t-1}$, defined as follows.

$$C_t^{B,J}(b_{t-J}^{t-1}) = \sup_{\pi_i^J(da_i|b_{i-J}^{i-1}) : i=t, t+1, \dots, n} \mathbf{E}^{\pi^J} \left\{ \sum_{i=t}^n \left[\int_{\mathbb{B}_i} \log \left(\frac{dQ_i(\cdot|b_{i-M}^{i-1}, A_i)}{d\mathbf{V}_i^{\pi^J}(\cdot|b_{i-J}^{i-1})}(b_i) \right) Q_i(db_i|b_{i-M}^{i-1}, A_i) - s\gamma_i^{B,K}(A_i, b_{i-K}^{i-1}) \right] \middle| B_{t-J}^{t-1} = b_{t-J}^{t-1} \right\} \quad (\text{III.163})$$

where $s \in [0, \infty)$ is the Lagrange and the term $(n+1)\kappa$ is not included.

Then the cost-to-go satisfies the following dynamic programming recursions.

$$C_n^{B,J}(b_{n-J}^{n-1}) = \sup_{\pi_n^J(da_n|b_{n-J}^{n-1})} \left\{ \int_{\mathbb{A}_n \times \mathbb{B}_n} \log \left(\frac{dQ_n(\cdot|b_{n-M}^{n-1}, a_n)}{d\nu_n^{\pi_n^J}(\cdot|b_{n-J}^{n-1})} (b_n) \right) Q_n(db_n|b_{n-M}^{n-1}, a_n) \otimes \pi_n^J(da_n|b_{n-J}^{n-1}) \right. \\ \left. - s \int_{\mathbb{A}_n} \gamma_n^{B,K}(a_n, b_{n-K}^{n-1}) \pi_n^J(da_n|b_{n-J}^{n-1}) \right\}, \quad (\text{III.164})$$

$$C_t^{B,J}(b_{t-J}^{t-1}) = \sup_{\pi_t^J(da_t|b_{t-J}^{t-1})} \left\{ \int_{\mathbb{A}_t \times \mathbb{B}_t} \log \left(\frac{dQ_t(\cdot|b_{t-M}^{t-1}, a_t)}{d\nu_t^{\pi_t^J}(\cdot|b_{t-J}^{t-1})} (b_t) \right) Q_t(db_t|b_{t-M}^{t-1}, a_t) \otimes \pi_t^J(da_t|b_{t-J}^{t-1}) \right. \\ \left. - s \int_{\mathbb{A}_t} \gamma_t^{B,K}(a_t, b_{t-K}^{t-1}) \pi_t^J(da_t|b_{t-J}^{t-1}) + \int_{\mathbb{A}_t \times \mathbb{B}_t} C_{t+1}^{B,J}(b_{t+1-J}^{t-1}) Q_t(db_t|b_{t-M}^{t-1}, a_t) \otimes \pi_t^J(da_t|b_{t-J}^{t-1}) \right\}, \quad t = n-1, \dots, 0. \quad (\text{III.165})$$

The characterization of the FTFI capacity is expressed via the $C_0^{B,J}(b_{-J}^{-1})$ and the fixed distribution $\mu_{B_{-J}^{-1}}(db_{-J}^{-1})$ by

$$C_{A^n \rightarrow B^n}^{FB,B,J}(\kappa) = \inf_{s \geq 0} \left\{ \int_{\mathbb{B}_{-J}^{-1}} C_0^{B,J}(b_{-J}^{-1}) \mu_{B_{-J}^{-1}}(db_{-J}^{-1}) - (n+1)s\kappa \right\}. \quad (\text{III.166})$$

It is obvious from the above recursions that, that the information structure, $\{B_{t-J}^{t-1} : t = 0, \dots, n\}$, of the control object, namely, $\{\pi_t^J(da_t|b_{t-J}^{t-1}) : t = 0, \dots, n\}$, induces transition probabilities of the controlled object, $\{\nu_t^{\pi_t^J}(db_t|b_{t-J}^{t-1}) : t = 0, \dots, n\}$ which are J -order Markov, resulting in a significant reduction in computational complexity of the above dynamic programming recursions. This is one of the fundamental differences, compared to other dynamic programming algorithms proposed in the literature, which do not investigate the impact of information structures, on the characterization of FTFI capacity, and by extension of feedback capacity.

Special Case-Unit Memory Channel Output (UMCO) $M = K = 1$. Since in this case, $J = 1$, the corresponding dynamic programming recursions are degenerate versions of (III.164), (III.165), obtained by setting $K = M = 1, J = 1$, i.e.,

$$Q_t(db_t|b_{t-M}^{t-1}, a_t) \mapsto Q_t(db_t|b_{t-1}, a_t), \quad \gamma_t^{B,K}(a_t, b_{t-K}^{t-1}) \mapsto \gamma_t^{B,1}(a_t, b_{t-1}), \\ \pi_t^J(da_t|b_{t-J}^{t-1}) \mapsto \pi_t^1(da_t|b_{t-1}), \quad C_t^{B,J}(b_{t-J}^{t-1}) \mapsto C_t^{B,1}(b_{t-1}), \quad t = n, \dots, 0. \quad (\text{III.167})$$

This degenerate dynamic programming recursion is the simplest, because the joint process $\{(A_i, B_i) : i = 0, \dots, n\}$ is jointly Markov (first-order), and the channel input conditional distribution is $\{\pi_i^1(da_i|b_{i-1}) : i = 0, \dots, n\}$. At each time t , the dynamic programming recursion involves 3-letters, $\{b_t, a_t, b_{t-1}\}$, where b_{t-1} is fixed, for $t = n, n-1, \dots, 0$.

It is noted that, for the case of finite alphabet spaces $\{(\mathbb{A}_i, \mathbb{B}_i) : i = 0, \dots, n\}$, the UMCO without transmission cost constraints is analyzed extensively by Chen and Berger in [30] (and it is discussed by Berger in [45]), under the assumption the optimal channel input conditional distribution satisfies conditional independence $P(da_i|a^{i-1}, b^{i-1}) = \pi_i^1(da_i|b_{i-1}), i = 0, \dots, n$, which then implies $\{(A_i, B_i) : i = 0, \dots, n\}$ is jointly Markov, and hence the corresponding characterization of FTFI capacity is given by $C_{A^n \rightarrow B^n}^{FB,B,1} = \sup_{\pi_i(da_i|b_{i-1}) : i=0, \dots, n} \sum_{i=0}^n I(A_i; B_i|B_{i-1})$. To the best of the authors knowledge, the current paper, provides, for the first time, a derivation of the fundamental assumptions, upon which the results derived in [30], are based on.

The main point to be made regarding this section, is that the information structure of the optimal channel input distribution maximizing directed information, can be obtained for many different classes of channels with memory, and many different classes of transmission cost functions, and that the corresponding conditional independence properties of optimal channel input distributions and characterizations of the FTFI capacity are generalizations of (I.10) and the two-letter capacity formulae of Shannon, corresponding to memoryless channels.

These structural properties of optimal channel input conditional distributions simplify the computation of the corresponding FTFI capacity characterization, and its per unit time limiting versions, the characterization of feedback capacity.

Remark III.4. (*Generalizations to channels with memory on past channel inputs*)

The reader may verify that the methodology developed this paper, to identify the information structures of optimal channel input distributions satisfying conditional independence, is also applicable to general channel distributions and transmission cost functions of the form,

$$\{\mathbf{P}_{B_i|B_{i-M}^{i-1}, A_{i-L}^i}(db_i|b_{i-M}^{i-1}, A_{i-L}^i) : i = 0, \dots, n\}, \quad \{\gamma_i(A_{i-N}^i, b_{i-K}^{i-1}) : i = 0, \dots, n\} \quad (\text{III.168})$$

($\{N, L\}$ are nonnegative integers) which depend on past source symbols. However, such generalizations of the structural properties of optimal channel input conditional distributions, are beyond the scope of this paper.

IV. APPLICATION EXAMPLE: GAUSSIAN LINEAR CHANNEL MODEL

The objective of the application example discussed below is to illustrate the role of the information structures of optimal channel distributions in deriving closed form expressions for feedback capacity, capacity without feedback, and to illustrate hidden aspects of feedback.

Consider the time-invariant version of a Gaussian-Linear Channel Model (G-LCM) of class B with transmission cost of Class B, defined by

$$B_i = C B_{i-1} + D A_i + V_i, \quad B_{-1} = b_{-1}, \quad i = 0, \dots, n, \quad (\text{IV.169})$$

$$\frac{1}{n+1} \sum_{i=0}^n \mathbf{E} \left\{ \langle A_i, R A_i \rangle + \langle B_{i-1}, Q B_{i-1} \rangle \right\} \leq \kappa, \quad R \in \mathbb{S}_{++}^{q \times q}, \quad Q \in \mathbb{S}_{+}^{p \times p} \quad (\text{IV.170})$$

$$\mathbf{P}_{V_i|V^{i-1}, A^i}(dv_i|v^{i-1}, a^i) = \mathbf{P}_{V_i}(dv_i) - a.a.(v^{i-1}, a^i), \quad V_i \sim N(0, K_{V_i}), \quad K_{V_i} = K_V \in \mathbb{S}_{++}^{p \times p}, \quad i = 0, \dots, n \quad (\text{IV.171})$$

where $\langle \cdot, \cdot \rangle$ denotes inner product of vectors, $\mathbb{S}_{++}^{q \times q}$ denotes the set of symmetric positive definite q by q matrices and $\mathbb{S}_{+}^{q \times q}$ the set of positive semi-definite matrices. The initial state b_{-1} is known to the encoder and decoder.

From Theorem III.3, the optimal channel input distribution maximizing directed information satisfies conditional independence $P_i(da_i|a^{i-1}, b^{i-1}) = \pi_i(da_i|b_{i-1}) - a.a.(a^{i-1}, b^{i-1}), i = 0, \dots, n$. Moreover, it can be easily shown, i.e., using the maximum entropy properties of Gaussian distributions, or by solving the corresponding dynamic programming recursion of the FTFI capacity, that the optimal distribution satisfying the average cost constraint is Gaussian, i.e., $\pi_i(da_i|b_{i-1}) \equiv \pi_i^g(da_i|b_{i-1}), i = 0, \dots, n$, which then implies the joint process is also Gaussian, i.e., $(A_i, B_i, V_i) \equiv (A_i^g, B_i^g, V_i), i = 0, \dots, n$, provided of course that the RV B_{-1} is also Gaussian.

Any such optimal channel input conditional distribution can be realized via an orthogonal decomposition as follows.

$$A_i^g \triangleq U_i^g + Z_i^g, \quad U_i^g = g_i^{B,1}(B_{i-1}^g) \equiv \Gamma_{i,i-1} B_{i-1}^g, \quad i = 0, \dots, n, \quad (\text{IV.172})$$

$$Z_i^g \text{ is independent of } (A^{g,i-1}, B^{g,i-1}), \quad Z^{g,i} \text{ is independent of } V^i, i = 0, \dots, n, \quad (\text{IV.173})$$

$$\{Z_i^g \sim N(0, K_{Z_i}) : i = 0, 1, \dots, n\} \text{ is an independent Gaussian process} \quad (\text{IV.174})$$

Moreover, substituting (IV.172) into (IV.169) the channel output process is given by

$$B_i^g = C B_{i-1}^g + D U_i^g + D Z_i^g + V_i, \quad i = 0, \dots, n. \quad (\text{IV.175})$$

The corresponding characterization of the FTFI capacity is the following.

$$C_{A^n \rightarrow B^n}^{FB,B,1}(\kappa) = \sup_{\{(g_i^{B,1}(\cdot), K_{Z_i}), i=0, \dots, n\} \in \mathcal{G}_{[0,n]}^{B,1}(\kappa)} \sum_{i=0}^n H(B_i^g|B_{i-1}^g) - H(V^n) \quad (\text{IV.176})$$

$$\mathcal{G}_{[0,n]}^{B,1}(\kappa) \triangleq \left\{ (g_i^{B,1}(b_{i-1}), K_{Z_i}), i = 0, \dots, n : \frac{1}{n+1} \mathbf{E}^{g^{B,1}} \left(\sum_{i=0}^n [\langle A_i^g, R A_i^g \rangle + \langle B_{i-1}^g, Q B_{i-1}^g \rangle] \right) \leq \kappa \right\}. \quad (\text{IV.177})$$

Suppose the pair $(g_i^{B,1}(\cdot), K_{Z_i}) \equiv (g^{B,1}(\cdot), K_Z), i = 0, \dots, n$, i.e., is restricted to time-invariant, and consider the per unit time limiting version of the characterization of FTFI capacity, defined by $C_{A^\infty \rightarrow B^\infty}^{FB}(\kappa) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n+1} \triangleq C_{A^n \rightarrow B^n}^{FB,B,1}(\kappa)$. Then one method to obtain $C_{A^\infty \rightarrow B^\infty}^{FB,B,1}(\kappa)$ is via dynamic programming as follows [32], [33].

Under appropriate conditions expressed in terms of the matrices $\{C, D, R, Q\}$ [33], there exists a pair $(J^{B.1,*}, C^{B.1}(b)), J^{B.1,*} \in \mathbb{R}$, $C^{B.1} : \mathbb{R}^p \mapsto \mathbb{R}$, which satisfies the following dynamic programming equation corresponding to $C_{A^\infty \rightarrow B^\infty}^{FB,B.1}(\kappa)$.

$$J^{B.1,*} + C^{B.1}(b) = \sup_{(u, K_Z) \in \mathbb{R}^q \times \mathbb{S}_+^{q \times q}} \left\{ \frac{1}{2} \log \frac{|DK_Z D^T + K_V|}{|K_V|} - \text{tr}(s R K_Z) + s \kappa - s [\langle u, Ru \rangle + \langle b, Qb \rangle] \right. \\ \left. + \mathbf{E}^{g^{B.1}} \left\{ C^{B.1}(B_0^g) \middle| B_{-1}^g = b \right\} \right\} \quad (\text{IV.178})$$

where $s \equiv s(\kappa) \geq 0$ is the Lagrange multiplier associated with the average transmission cost constraint.

It can be verified that the solution to the dynamic programming is given by

$$C^{B.1}(b) = -\langle b, Pb \rangle, \quad (\text{IV.179})$$

$$J^{B.1,*} = \sup_{K_Z \in \mathbb{S}_+^{q \times q}} \left\{ \frac{1}{2} \log \frac{|DK_Z D^T + K_V|}{|K_V|} + s \kappa - \text{tr}(s R K_Z) - \text{tr}(P [DK_Z D^T + K_V]) \right\} \quad (\text{IV.180})$$

and that the optimal stationary policy $g^{B.1,*}(\cdot)$ is given by

$$g^{B.1,*}(b) = \Gamma^* b, \quad (\text{IV.181})$$

$$\Gamma^* = -\left(D^T P D + s R\right)^{-1} D^T P C, \quad (\text{IV.182})$$

$$P = C^T P C + s Q - C^T P D \left(D^T P D + s R\right)^{-1} \left(C^T P D\right)^T. \quad (\text{IV.183})$$

$$\text{spec}(C + D \Gamma^*) \subset \mathbb{D}_o. \quad (\text{IV.184})$$

where $\text{spec}(A) \subset \mathbb{C}$ denotes the spectrum of a matrix $A \in \mathbb{R}^{q \times q}$, i.e., the set of all its eigenvalues, and $\mathbb{D}_o \triangleq \{c \in \mathbb{C} : |c| < 1\}$ denotes the open unit disc of the space of complex number \mathbb{C} . Note that (IV.183) is the well-known Riccati equation of Gaussian Linear Quadratic stochastic optimal control problems, and $\Gamma^* \equiv \Gamma^*(P)$ corresponds to the positive semi-definite solution $P \succeq 0$ of the Riccati equation [33], satisfying (IV.184), to ensure the eigenvalues of the closed loop channel output recursion (IV.175), i.e., corresponding to $U_i^{g,*} = \Gamma^* B_{i-1}^g, i = 0, \dots$, are within the open unit disc \mathbb{D}_o .

The optimal covariance K_Z^* is determined from the optimization problem (IV.180) and the Lagrange multiplier, for a given κ , i.e., $s \equiv s(\kappa)$ is found from the average constraint. The feedback capacity is given by

$$C_{A^\infty \rightarrow B^\infty}^{FB,B.1}(\kappa) = J^{B.1,*} \equiv J^{B.1,*}(\kappa), \quad \kappa \in [\kappa_{\min}, \infty) \subset [0, \infty). \quad (\text{IV.185})$$

The analysis of the Multiple Input Multiple Output G-LCM is done in [46], and requires extensive investigation of properties of solutions to Riccati equations. The complete solution for the scalar G-LCM is presented below, to illustrate additional features, which are not given in [46].

Scalar Case, $p = q = 1, D = 1$. By solving (IV.183), the positive semi-definite solution of the Riccati equations is given by

$$P = \frac{s(Q - R + C^2 R + F)}{2}, \quad F = \sqrt{(R[C - 1]^2 + Q)(R[C + 1]^2 + Q)} \quad (\text{IV.186})$$

The value of $\Gamma^* \equiv \Gamma^*(P)$ is obtained by substituting the positive semi-definite solution of the Riccati equation in (IV.182), to obtain

$$\Gamma^* = -\frac{C(Q - R + C^2 R + F)}{Q + R + C^2 R + F}, \quad |C + \Gamma^*| < 1. \quad (\text{IV.187})$$

This is valid irrespectively of whether C is stable, i.e., $|C| < 1$ or unstable, i.e., $|C| \geq 1$, and includes, as we show shortly, the special case $Q = C = 0$, i.e., corresponding to the memoryless channel.

The optimal covariance K_Z^* , is obtained by solving the optimization problem (IV.180), which gives

$$K_Z^* = \frac{1}{s(Q + R + C^2 R + F)} - K_V \geq 0 \quad (\text{IV.188})$$

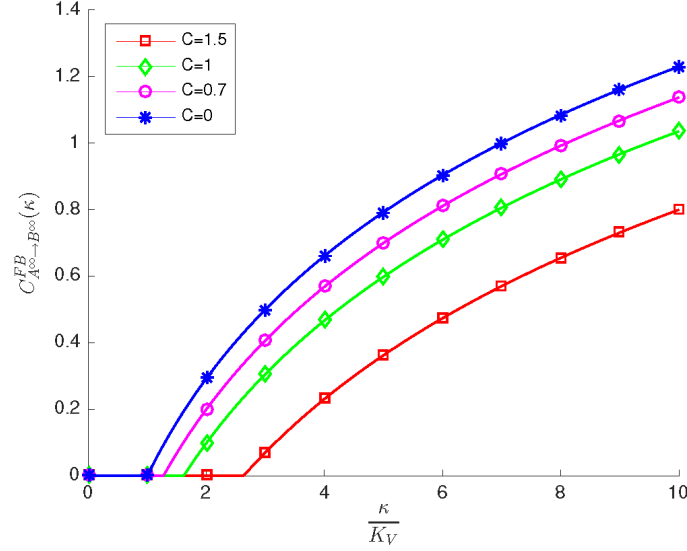


Fig. IV.2. Feedback capacity for the scalar Linear-Gaussian Channel Model ($Q = 1, R = 1, D = 1$).

while the Lagrange multiplier, s , is found from the average constraint or by performing the infimum over $s \geq 0$ of $J^{B.1,*}$ evaluated at (P, K_Z^*) given by (IV.180), to obtain

$$s \equiv s(\kappa) = \frac{1}{2(\kappa + K_V R)}. \quad (\text{IV.189})$$

The minimum power κ required so that the optimal covariance is non-negative, i.e., $K_Z^* \geq 0$ is found by substituting (IV.189) in (IV.188) and it is given by

$$\kappa_{min} = \frac{K_V (Q - R + C^2 R + F)}{2} \geq 0. \quad (\text{IV.190})$$

Finally by substituting (IV.188) and (IV.189) in (IV.180), the following expression of the feedback capacity is obtained.

$$C_{A^\infty \rightarrow B^\infty}^{FB,B.1}(\kappa) = \begin{cases} 0 & \text{if } \kappa \in [0, \kappa_{min}) \\ \frac{1}{2} \log \left(\frac{2(\kappa + K_V R)}{K_V (Q + R + C^2 R + F)} \right) & \text{if } \kappa \in [\kappa_{min}, \infty). \end{cases} \quad (\text{IV.191})$$

It can be verified that the value of κ_{min} depends on the values of $Q = 0$, $Q > 0$ and $|C| < 1$, $|C| \geq 1$. For $Q = C = 0$ the feedback capacity $C_{A^\infty \rightarrow B^\infty}^{FB,B.1}(\kappa)$ degenerates to that of memoryless channels, as expected.

Next, specific special cases are analyzed to gain additional insight on the dependence of capacity on $|C| < 1$ and $|C| \geq 1$.

(a) Suppose $Q = R = 1$. Then

$$F = \sqrt{C^4 + 4}, \quad s = \frac{1}{2(\kappa + K_V)}, \quad K_Z^* = \frac{2\kappa - K_V \sqrt{C^4 + 4} - C^2 K_V}{\sqrt{C^4 + 4} + C^2 + 2}. \quad (\text{IV.192})$$

The feedback capacity is given by

$$C_{A^\infty \rightarrow B^\infty}^{FB,B.1}(\kappa) = \begin{cases} 0 & \text{if } \kappa \in [0, \kappa_{min}) \\ \frac{1}{2} \log \left(\frac{2(\kappa + K_V)}{K_V (\sqrt{C^4 + 4} + C^2 + 2)} \right) & \text{if } \kappa \in [\kappa_{min}, \infty) \end{cases} \quad (\text{IV.193})$$

where $\kappa_{min} = \frac{K_V (\sqrt{C^4 + 4} + C^2)}{2}$. Clearly, if $C = 0$ then $\kappa_{min} = K_V$ and this is attributed to the fact that, the power transfer of the channel output process is reflected in the average power constraint, i.e., $Q = 1$.

The feedback capacity for $D = Q = R = 1$ is illustrated in Fig. IV.2, for stable and unstable values of the parameter C . It

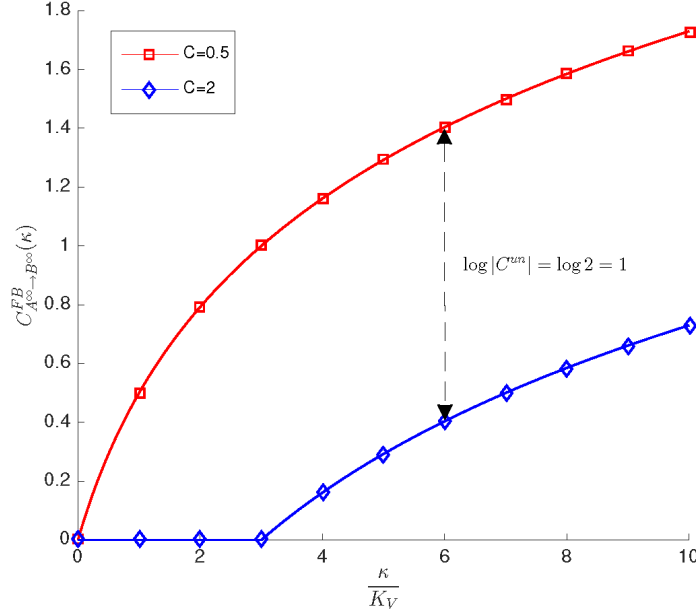


Fig. IV.3. Feedback capacity for the scalar Linear-Gaussian Channel Model ($Q=0, R=1, D=1$), where C^{un} denotes the value of C of the unstable channel ($C=2$).

illustrates that there is a minimum value $\kappa_{min} > 0$, because the transmission cost function includes the power of the channel output process, and because of this, part of the power κ is transfer to the channel output.

(b) Suppose $D=R=1, Q=0$. Then the cost constraint is independent of past channel output symbols, and $F=C^2-1$ which yields

$$P = \begin{cases} 0 & \text{if } |C| < 1 \\ C^2 - 1 & \text{if } |C| \geq 1. \end{cases} \quad (\text{IV.194})$$

The optimal strategy which achieves feedback capacity $C_{A^\infty \rightarrow B^\infty}^{FB,B.1}(\kappa)$ is given by

$$(\Gamma^*, K_Z^*) = \begin{cases} (0, \kappa), & \kappa \in [0, \infty) & \text{if } |C| < 1 \\ \left(-\frac{C^2-1}{C}, \frac{\kappa + K_V(1-C^2)}{C^2} \right), & \kappa \in [\kappa_{min}, \infty), \quad \kappa_{min} \triangleq (C^2-1)K_V & \text{if } |C| \geq 1 \\ \left(-\frac{C^2-1}{C}, 0 \right), & \kappa \in [0, \kappa_{min}], & \text{if } |C| \geq 1. \end{cases} \quad (\text{IV.195})$$

Let $C_{A^\infty \rightarrow B^\infty}^{FB,Stable}(\kappa)$ denote the feedback capacity if the channel is stable, i.e., $|C| < 1$ and $C_{A^\infty \rightarrow B^\infty}^{FB,Unstable}(\kappa)$ denote the feedback capacity if the channel is unstable, i.e., $|C| \geq 1$. Then, the corresponding feedback capacity is given by the following expressions.

For $|C| < 1$:

$$C_{A^\infty \rightarrow B^\infty}^{FB,B.1}(\kappa) \triangleq C_{A^\infty \rightarrow B^\infty}^{FB,Stable}(\kappa) = \frac{1}{2} \log \left(1 + \frac{\kappa}{K_V} \right), \quad \kappa \in [0, \infty). \quad (\text{IV.196})$$

For $|C| \geq 1$:

$$C_{A^\infty \rightarrow B^\infty}^{FB,B.1}(\kappa) \triangleq C_{A^\infty \rightarrow B^\infty}^{FB,Unstable}(\kappa) = \begin{cases} \frac{1}{2} \log \left(1 + \frac{\kappa}{K_V} \right) - \log |C| & \text{if } \kappa \in [\kappa_{min}, \infty) \\ 0 & \text{if } \kappa \in [0, \kappa_{min}]. \end{cases} \quad (\text{IV.197})$$

Then it is clear from (IV.196) and (IV.197), that

$$C_{A^\infty \rightarrow B^\infty}^{FB, Unstable}(\kappa) = C_{A^\infty \rightarrow B^\infty}^{FB, Stable}(\kappa) - \log |C|, \quad \kappa \in [\kappa_{min}, \infty). \quad (IV.198)$$

Therefore, the rate loss due to the instability of the channel is given by

$$\text{Rate Loss of Unstable Channels} \triangleq C_{A^\infty \rightarrow B^\infty}^{FB, Stable}(\kappa) - C_{A^\infty \rightarrow B^\infty}^{FB, Unstable}(\kappa) = \begin{cases} \frac{1}{2} \log \left(1 + \frac{\kappa}{K_V} \right), & \kappa \in [0, \kappa_{min}] \\ \log |C|, & \kappa \in [\kappa_{min}, \infty). \end{cases} \quad (IV.199)$$

The feedback capacity of a stable channel ($C = 0.5$) and an unstable channel ($C = 2$), is depicted in Fig. IV.3. The dotted arrow denotes the rate loss due to the instability of the channel with parameter $C = 2$, which is equal to $\log 2 = 1$ bit. It is worth noting that for unstable channels, the feedback capacity is zero, unless the power κ exceeds the critical level κ_{min} . This the minimum power required to stabilize the channel. Beyond this threshold all the remaining power ($\kappa - \kappa_{min}$) is allocated to information transfer.

On the other hand, if the channel is stable, i.e., $|C| < 1$, since $Q = 0$, there is no emphasis on power transfer of the channel output process, and feedback capacity degenerates to the capacity of memoryless additive Gaussian noise channel, i.e., feedback does not increase capacity compared to that of memoryless channels, as verified from (IV.196). This is, however, fundamentally different from the case $|C| < 1$ and $Q > 0$ discussed in (a).

V. CONCLUSION

Stochastic optimal control theory and a variational equality of directed information are applied, to develop a methodology to identify the information structures of optimal channel input conditional distributions, which maximize directed information, for certain classes of channel conditional distributions and transmission cost constraints. The subsets of the maximizing distributions are characterized by conditional independence.

One of the main theorems of this paper states that, for any channel conditional distribution with finite memory on past channel outputs, subject to any average transmission cost constraint corresponding to a specific transmission cost function, the information structure of the optimal channel input conditional distribution, which maximizes directed information, is determined by the maximum of the memory of the channel distribution and the functional dependence of the transmission cost function on past channel outputs. This theorem provides, for the first time, a direct analogy, in terms of the conditional independence properties of maximizing distributions, between the characterization of feedback capacity of channels with memory, and Shannon's two-letter characterization of capacity of memoryless channels.

Whether a similar method, based on stochastic optimal control theory and variational equalities of mutual and directed information, can be developed for extremum problems of capacity of channels with memory and without feedback, and for general extremum problems of network information theory, remains, however to be seen.

REFERENCES

- [1] H. Marko, "The bidirectional communication theory—A generalization of information theory," *IEEE Transactions on Communications*, vol. 21, no. 12, pp. 1345–1351, Dec. 1973.
- [2] J. L. Massey, "Causality, feedback and directed information," in *International Symposium on Information Theory and its Applications (ISITA '90)*, Nov. 27–30 1990, pp. 303–305.
- [3] H. Permuter, T. Weissman, and A. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 2, pp. 644–662, Feb. 2009.
- [4] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Transactions on Information Theory*, vol. 49, no. 1, pp. 4–21, Jan. 2003.
- [5] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.

- [6] R. L. Dobrushin, "General formulation of Shannon's main theorem of information theory," *Usp. Math. Nauk.*, vol. 14, pp. 3–104, 1959, translated in *Am. Math. Soc. Trans.*, 33:323–438.
- [7] M. Pinsker, *Information and Information Stability of Random Variables and Processes*. Holden-Day Inc, San Francisco, 1964, translated by Amiel Feinstein.
- [8] R. T. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., New York, 1968.
- [9] R. E. Blahut, *Principles and Practice of Information Theory*, ser. in Electrical and Computer Engineering. Reading, MA: Addison-Wesley Publishing Company, 1987.
- [10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.
- [11] S. Ihara, *Information theory for Continuous Systems*. World Scientific, 1993.
- [12] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1147–1157, July 1994.
- [13] T. S. Han, *Information-Spectrum Methods in Information Theory*, 2nd ed. Springer-Verlag, Berlin, Heidelberg, New York, 2003.
- [14] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, Swiss Federal Institute of Technology (ETH), December 1998.
- [15] E. A. Gamal and H. Y. Kim, *Network Information Theory*. Cambridge University Press, December 2011.
- [16] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech.*, vol. 27, pp. 379–423, July 1948.
- [17] R. L. Dobrushin, "Information transmission in a channel with feedback," *Theory of Probability and its Applications*, vol. 3, no. 2, pp. 367–383, 1958.
- [18] T. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Transactions on Information Theory*, vol. 35, no. 1, pp. 37–43, Jan. 1989.
- [19] C. D. Charalambous and C. Kourtellis, "Structural properties of capacity achieving information lossless randomized encoders for feedback channels with memory and transmission cost-part ii," *IEEE Transactions on Information Theory*, 2015, submitted, November 2015.
- [20] Y.-H. Kim, "Feedback capacity of stationary Gaussian channels," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 57–85, 2010.
- [21] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback-I: no bandwidth constraints," *IEEE Transactions on Information Theory*, vol. 12, April 1966.
- [22] S. Yang, A. Kavcic, and S. Tatikonda, "On feedback capacity of power-constrained Gaussian noise channels with memory," *Information Theory, IEEE Transactions on*, vol. 53, no. 3, pp. 929–954, March 2007.
- [23] S. Butman, "A general formulation of linear feedback communications systems with solutions," *IEEE Transactions on Information Theory*, 1969.
- [24] —, "Linear feedback rate bounds for regressive channels," *IEEE Transactions on Information Theory*, 1976.
- [25] H. Permuter, P. Cuff, B. Van Roy, and T. Weissman, "Capacity of the trapdoor channel with feedback," *IEEE Transactions on Information Theory*, 2010.
- [26] O. Elishco and H. Permuter, "Capacity and coding of the ising channel with feedback," *IEEE Transactions on Information Theory*, vol. 60, no. 9, pp. 3138–3149, June 2014.
- [27] H. Permuter, H. Asnani, and T. Weissman, "Capacity of a post channel with and without feedback," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6041–6057, Oct 2014.
- [28] C. Kourtellis and C. Charalambous, "Capacity of binary state symmetric channel with and without feedback and transmission cost," in *IEEE Information Theory Workshop (ITW)*, May 2015.
- [29] S. Yang, A. Kavcic, and S. Tatikonda, "Feedback capacity of finite-state machine channels," *Information Theory, IEEE Transactions on*, vol. 51, no. 3, pp. 799–810, March 2005.
- [30] J. Chen and T. Berger, "The capacity of finite-state Markov channels with feedback," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 780–798, March 2005.
- [31] P. E. Caines, *Linear Stochastic Systems*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York, 1988.
- [32] O. Hernandez-Lerma and J. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, ser. Applications of Mathematics Stochastic Modelling and Applied Probability. Springer Verlag, 1996, no. v. 1.
- [33] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice Hall, 1986.
- [34] P. Stavrou, C. Charalambous, and C. Kourtellis, "Sequential necessary and sufficient conditions for optimal channel input distributions of channels with memory and feedback," in *IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain*, July 10–15 2016.
- [35] C. Kourtellis, "Nonanticipative information theory," Ph.D. dissertation, University of Cyprus (UCY), Nicosia, Cyprus, March 2014.
- [36] C. Kourtellis, C. Charalambous, and J. Boutros, "Nonanticipative transmission of sources and channels with memory," in *IEEE International Symposium on Information Theory (ISIT)*, 2015.

- [37] C. P. Stavrou and C. Kourtellaris, "Sequential necessary and sufficient conditions for capacity achieving distributions of channels with memory and feedback," *IEEE Transactions on Information Theory*, 2016, submitted, April 2016 (available on ArXiv).
- [38] C. D. Charalambous and P. A. Stavrou, "Directed information on abstract spaces: Properties and variational equalities," *submitted to IEEE Transactions on Information Theory*, 2013. [Online]. Available: <http://arxiv.org/abs/1302.3971>
- [39] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, July 1972.
- [40] N. U. Ahmed and C. D. Charalambous, "Stochastic minimum principle for partially observed systems subject to continuous and jump diffusion processes and driven by relaxed controls," *SIAM Journal on Control and Optimization*, vol. 51, no. 4, pp. 3235–3257, 2013.
- [41] C. D. Charalambous and R. J. Elliott, "Classes of nonlinear partially observable stochastic optimal control problems with explicit optimal control laws," *SIAM Journal on Control and Optimization*, vol. 36, no. 2, pp. 542–578, 1998.
- [42] J. H. van Schuppen, *Mathematical Control and System Theory of Discrete-Time Stochastic Systems*. Preprint, 2010.
- [43] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Volume 1*. Athena Scientific, 1995.
- [44] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York, 1969.
- [45] T. Berger, "Living Information Theory," *IEEE Information Theory Society Newsletter*, vol. 53, no. 1, March 2003.
- [46] C. D. Charalambous, C. K. Kourtellaris, and S. Loyka, "Capacity achieving distributions & information lossless randomized strategies for feedback channels with memory: The LQG theory of directed information-part II," *IEEE Transactions on Information Theory*, submitted, April 2016. [Online]. Available: <http://arxiv.org/abs/1604.01056>